

Ripe opportunities for regional collaboration in east Asia

East Asian collaboration in molecular biology could enhance the scientific viability of the region, but it requires more enthusiastic support from potential participants — particularly Japan.

Despite the flowering of scientific excellence in many east Asian countries in recent years, researchers in that part of the world in search of collaborators are often inclined to look right past each other to colleagues in the United States or Europe. The tendency persists despite the geographical convenience of local collaboration, the common scientific and cultural heritage of the region, and the knowledge that major institutions in the United States, in particular, are liable to treat collaborators from abroad as junior partners.

Most countries in east Asia regard themselves as being short of top-level researchers and, like Singapore and Hong Kong, often rely on attracting such researchers from abroad to fill such positions (see News Feature, page 370, and Naturejobs, page 4). But if these nations are to attract such people — and to keep their own best talent — they need to work far more closely with their neighbours.

Since 1997, researchers in the biological sciences have been trying to achieve this through an informal research network, known as the Asia-Pacific International Molecular Biology Network (IMBN). The IMBN's 250 members want to establish a regional equivalent of the European Molecular Biology Organization (EMBO), with its own laboratory and all the trappings of an international, collaborative research institution. But the concept has so far made limited headway, partly because of lack of support from the region's strongest nations.

The kind of scientific collaboration that IMBN members have in mind could proceed now, without waiting for closer political ties

between participant nations. Both EMBO and CERN, the European particle physics laboratory, were set up without relying on the European Union or its predecessor organizations.

The key challenge for IMBN is to obtain commitments from the strongest nations in the region — Japan, Australia and Singapore — in order to build the project's momentum. Global experience suggests that collaboration cannot occur without top-level political commitment from national governments. Officials in Singapore say that it may be prepared to match any funds forthcoming from Japan, although the Japanese government has yet to fully embrace the concept.

But Japan, arguably, has more to gain from the project than anybody else. By supporting it, Japan could show regional leadership at low diplomatic risk and relatively little financial cost. The project would also help Japan to cultivate the international outlook that its scientific community has struggled to achieve. Additionally, an efficient Asian molecular-biology laboratory could provide a badly needed voice for the region's biologists. They may otherwise find themselves unheard in international discussions, where their European colleagues are represented by EMBO and their American ones by the National Institutes of Health.

In Europe, CERN and EMBO have shown the importance of formal regional cooperation, particularly in providing a counterweight to the United States. East Asia could make similar gains by actively exploring the options for regional collaboration in molecular biology. ■

Back from the brink

The government's threat to suspend clinical trials at America's largest medical school highlights an impasse over funding.

Biomedical researchers in the United States were genuinely stunned last week to learn that the federal government was suspending the clinical research programmes at Johns Hopkins University (JHU) in Baltimore, the largest academic medical centre in the country, and one of the most revered.

The suspension announced by the fledgling Office for Human Research Protections (OHRP) was condemned in unusually forthright terms by JHU researchers and administrators, who clearly felt that the government had overreached itself. Dispensing with the usual pleasantries, the university went on the offensive, publicly condemning what it regards as the OHRP's rushed judgement.

Both sides are now backing off somewhat (see page 363). Government spokesmen are explaining, rather meekly, that trials will be allowed to continue if it is clearly in patients' interests that they do so. Johns Hopkins has negotiated an action plan that will satisfy the OHRP's concerns and allow the suspension to be gradually lifted.

However, the charges being made against JHU's system for the internal review of clinical research protocols are not trivial. It has

been widely noted over the past several years that the institutional review boards (IRBs) that do this job across the United States are badly over-stretched. It is surprising, to put it mildly, that JHU obtained approval for the non-therapeutic administration to healthy volunteers of an asthma-inducing agent whose toxicity could have been verified by a simple Medline search. And it is salutary that the death of one of those volunteers, 24-year-old laboratory technician Ellen Roche, was first made public through the good offices of *The Baltimore Sun*, rather than through official channels.

Clinical researchers confront illness and death every day and are understandably impatient with bureaucratic requirements imposed by agencies such as the OHRP to cater for sometimes-hypothetical public concerns. The work of the researchers is indeed too valuable to halt. But if the threat of such a suspension at an establishment of Johns Hopkins' standing helps to dispel complacency about IRB reform — and forces the medical schools and the Congress to stop passing the buck and decide who should pay for a properly resourced IRB system — then it will have served a useful purpose. ■





Fast treatment
Europe may halve
its approval time
for new drugs
p364



Time for change
Talks wrapped up in
Bonn as Kyoto deal
moves forward
p365



Particle puzzle
NASA hopes to find
out how the Solar
System was born
p366



Bang on target
Physicists pledge to
develop large linear
particle collider
p367

Johns Hopkins researchers fume over government crackdown

Meredith Wadman, Washington

A government-ordered suspension of clinical trials at the largest academic medical centre in the United States has provoked outrage among administrators and investigators.

The shutdown on 19 July at Johns Hopkins University (JHU) in Baltimore was partially revoked after only three days. But its impact will continue for months, as some 2,200 research protocols will be reviewed by ethics boards before they can recommence.

The shutdown followed the death on 2 June of a previously healthy, 24-year-old volunteer in an asthma experiment in which she inhaled hexamethonium, a known lung toxin (see *Nature* **411**, 873; 2001). The university accepted "full responsibility" for her death in a statement accompanying a 16 July report on the incident.

The suspension of research at JHU — which the government partially lifted on 22 July after the university submitted an action plan to correct lapses in ethical monitoring — brought tensions between US clinical investigators and government regulators to a head.

The university, in a press release, called the action "unwarranted, unnecessary, paralyzing and precipitous", and said that the government had acted "in utter disregard of patients' health and potentially of life".

"It's almost paradoxical in my mind that an organization whose responsibility is for the safety of patients could by their action put so many patients in my hospital in harm's way," says George Dover, chief paediatrician at the Johns Hopkins Children's Center, part of the university's hospital.

Patient protection

The US Department of Health and Human Services rejected this criticism, pointing out that the shutdown exempted trials in which it was in the best interests of patients to continue. "We don't want to endanger people's lives. Our goal is to protect people's lives," says Bill Hall, a spokesman.

The shutdown pleased advocates of tougher protection for human subjects. "There is a culture now in academic research that needs to be really shaken," says Vera Hassner Sharav, president of the New York-based

Alliance for Human Research Protection. "A death occurred because of non-compliance," she contends, "and that's not a good enough reason to apply sanctions?"

In recent years, the government has enforced the rules that protect research subjects more strictly, shutting down research at several institutions including the University of California at Los Angeles and Duke University in North Carolina. But JHU, with over \$300 million in federal funding and more than 137,000 patients enrolled in clinical trials, is its biggest target yet. The suspension also tested the mettle of the government's new Office for Human Research Protections (OHRP), which ordered it. The OHRP was created last year to take over the watchdog role previously performed by a small office inside the National Institutes of Health, the biomedical research agency that funds most of the research. It reports directly to Tommy Thompson, the health secretary.

Minutes missing

On 22 July, the OHRP announced that it would allow about 500 protocols at JHU to resume immediately. These protocols either involve minimal risk or were adequately documented as having been substantially reviewed in recent minutes of JHU's Institutional Review Boards (IRBs). One of the OHRP's complaints with JHU is that minutes do not exist for 18 of the past 21 IRB meetings, dating back to October 2000.

The remaining 2,200 protocols must be re-reviewed by one of JHU's three IRBs, which administrators estimate will take 400 hours of committee time, lasting weeks or months.

Although the university is "pleased" with the formal end to the shutdown, the OHRP has "imposed a whole new set of paperwork requirements that can only impede the resumption of valuable scientific research", says Gary Stephenson, a JHU spokesman.

Even before the suspension was lifted, the university was continuing with hundreds of trials in which it judged that the best interests of patients would be hurt by stopping.

But clinical investigators at JHU were frustrated and angry at what they saw as an overblown reaction to an isolated incident.



Hospital blanket: the scale of the shutdown has angered clinical researchers at Johns Hopkins.

"It seems unnecessarily harsh and broad to shut down the whole system," says Bruce Bochner, an allergist who has had to halt studies of white blood cells that involve drawing peripheral blood from patients with hay fever and asthma.

Reza Shadmehr, an associate professor of biomedical engineering and neuroscience, analyses motor-control deficits in volunteers with brain diseases, who move robotic arms in his lab. He lost four days of data when his experiments were shut down. But he says that he is more concerned about the long-term impact of the suspension on volunteer trust. "This may impede the ability of at least this scientist to get the trust that he used to have," he says. "I think it's going to take us a few years to rebuild that."

MAXWELL BOAM

Drugs approval process gets speed treatment

Erica Klarreich, London

The European Union (EU) is mulling over changes that could halve the average time taken to approve new drug treatments.

A proposal adopted by the European Commission (EC) on 18 July would, its supporters say, cut the average licensing time for new active substances from 18 months to about 9 months. "We want to increase the availability of new and innovative medicines on the European market," says Erkki Liikanen, the EC commissioner for enterprise.

The plan will now go to the Council of Ministers, which represents the 15 members of the EU, and to the European Parliament. Both can suggest amendments. Before it can be implemented, the plan must be approved by the council and the parliament — a process that may take several years and modify the proposals significantly.

EC officials say the new plan would make the drug approvals process in Europe faster than that in the United States — although Liikanen denies that the EC is in a race with the US Food and Drug Administration (FDA) to speed up approval. "We just want to guarantee that European patients and industry have the same opportunities as anywhere else," he says.

The plan would shorten deadlines for various parts of the approval process, introducing a 'fast track' for especially promising



No safety compromise: Liikanen says the short-cuts would be bureaucratic, not scientific.

medicines. To streamline the process, the London-based European Agency for the Evaluation of Medicinal Products would be given much broader responsibilities. At present, drug companies can choose to go through this agency to obtain EU-wide approval for drugs containing new active substances. However, most medicines are still licensed by individual member states.

In the United States, a recent acceleration of the FDA's approval processes has been criticized by consumer groups. And some health experts warn that speeding up the European procedure may endanger patients. "Probably what will happen is that the amount of data that regulators check independently will be reduced to make deadlines," says John Abraham, director of the

Centre for Research in Health and Medicine at the University of Sussex.

A faster approval process will require the commission to spend more on licensing, draining resources from other public-health needs, says Joe Collier, a health-policy specialist at St George's Hospital Medical School in London. "The commercial arm of the European Commission has driven these changes, not the public-health arm," he says.

But Liikanen says the changes would not compromise drug safety. "The cuts will be on the bureaucratic side, not the scientific side," he says.

The pharmaceutical industry gave the plan a cautious endorsement. "Broadly speaking, we welcome efforts to approve drugs more quickly," says Richard Ley, a spokesman for the Association of the British Pharmaceutical Industry.

The commission thinks that the slow pace of approvals in Europe has driven away pharmaceutical research in recent years. Europe's share of the global pharmaceutical market has dwindled from 32% to 22% in the past decade, whereas the US share has grown from 31% to 43%, according to the European Federation of Pharmaceutical Industries and Associations. But some observers say that price controls in Europe — and their absence in the United States — has provided most of the impetus for the change. ■

AP/TIMOTHY CHARLIER

Scientists fear new guidelines will stifle basic research

David Cyranoski, Tokyo

Science-policy guidelines issued by Japanese prime minister Junichiro Koizumi have ignited a fire-storm of protest among the country's leading researchers.

The scientists are upset because they say the guidelines, which strongly emphasize the economic value of research and development, don't provide strong enough support for basic scientific research.

The guidelines were issued on 11 July by the Council for Science and Technology Policy, the top science-policy-making body in the government, which Koizumi chairs. They call for a strategic focus on four key areas of research — life sciences, information technology, environmental research and nanotechnology — and the reform of Japan's science and technology system to build industrial competitiveness, a vigorous economy, better health care and a safe environment.

But 18 former and existing heads of some of the country's leading research institutes wrote an open letter to Koizumi, stating that the guidelines will stifle basic



Koizumi: faced with protests.

research by "completely mobilizing science and technology to meet short-term goals aimed primarily at industrial competitiveness".

According to Yoshiki Hotta, director of the National Institute of Genetics, who helped to draft the letter, researchers outside the specified fields will either lose support or "do their work under the

cover of some other project that fits the government specifications. There's no clear allocation for basic science in the guidelines, and this is dangerous," he says.

A senior researcher at Tokyo University's medical school, who declined to be named, added that the guidelines "will lead to demoralization of scientists, as they keep running after whatever goal is set by the government".

The letter earned its authors a 15-minute audience with Koji Omi, a cabinet-level

minister for science and technology policy, who told them that the guidelines make sufficient allowance for basic research. The guidelines state that Japan should "achieve international-class, high-quality basic research that will clear a path to the future". But critics say such words merely pay lip-service to the problem.

Hotta says that the group wrote the open letter — a dramatic gesture by Japanese standards — because there is no proper channel for scientists to give their opinions to the government. Many researchers criticize the Council for Science and Technology Policy for failing to represent their views. The council is composed mostly of ministers and industry representatives, and researchers say that the three academic scientists on it have limited influence.

Some industrialists echo the scientists' complaints. The government has "no clear understanding of what the best research is — and yet they want to move on to application," says Toshiaki Ikoma, president of Texas Instruments in Japan. ■

MICHAEL EVANS/GETTY IMAGES

'Political fix' saves Kyoto deal from collapse

Jim Giles, Bonn

The Kyoto Protocol took a critical, if shaky, step forward this week as the latest round of climate-change talks ended in Bonn, Germany. Almost 180 countries signed up to rules that take the agreement on limiting greenhouse-gas emissions a step closer to coming into force next year.

The talks were marked by a series of concessions by the European Union (EU). The United States' withdrawal from the accord earlier this year, together the failure of the previous set of talks, held in The Hague last November, had raised fears that the process could collapse altogether. The compromises agreed on 23 July, after days of round-the-clock negotiations, were accepted by the EU to prevent countries such as Japan and Canada following the United States in rejecting the protocol.

The main EU concession was a climb-down over the use of carbon sinks — projects such as planting new forests to absorb carbon dioxide. The original protocol allowed a range of sink activities to be used as credits to increase the amount of greenhouse gases that countries are entitled to emit. But the Umbrella Group — a coalition of industrialized countries that includes Japan and Canada — had been pushing for the inclusion of further sink projects, such as changing the way in which existing forests are managed.

With delegates desperate for agreement, the Umbrella Group succeeded in winning a better deal than the one the EU had rejected in The Hague, allowing them to gain credit from forest-management projects and other new sinks. The EU also had to back down over plans to put a cap on the total number of credits that countries can earn from sinks and projects such as investing in emissions-reducing technology in developing countries. And although limits were agreed on the extent of the newly included sinks, Japan and Canada negotiated exemptions from these until 2010. "It's a political fix," says a British member of the EU negotiating team. "That's the price of the deal."

The final hours of the meeting also saw exhausted delegates defer a decision on the penalties for countries that fail to meet their targets, which the EU wants to be legally binding. But the EU did secure one of its aims — the rejection of nuclear power as a means of cutting emissions. Umbrella Group countries had wanted to be able to claim credits by replacing fossil-fuel power stations with nuclear ones. Some had also requested credit for helping developing countries to build nuclear power stations.

Environmental groups are disappointed that the treaty has been watered down. According to the WWF, the current agreement will achieve a 2% cut in greenhouse-gas emissions by 2010, rather than the 5%

predicted under the original protocol. Sink projects are likely to improve this figure somewhat, although it is unclear by how much.

Nevertheless, the main reaction is one of

relief that the Kyoto process is not dead. "This will send a message to the market," says Jennifer Morgan of the WWF. "It must factor in the cost of carbon."



Real deal: delegates in Bonn have settled on a watered-down version of the original agreement.

US rejects bioweapon inspections

Jonathan Knight, San Francisco

A plan that would give teeth to the Biological Weapons Convention seems to be doomed. US negotiators will tell this week's talks on the convention in Geneva that they still strongly object to the inclusion of a verification procedure in the treaty.

In all, 140 countries have ratified the convention since it was hammered out almost 30 years ago. But the treaty contains no provision for verification, a loophole that allowed the Soviet Union to operate dozens of germ-warfare facilities in the 1970s and 80s.

Attempts to develop a verification plan began in 1995. It was hoped that the latest draft, released in March, would address the concerns of many participants, including the United States. But as *Nature* went to press, US representatives were set to announce that they would not sign it in its current form.



Behind the mask: the United States is wary of revealing the secrets of its biological defences.

The United States has said in the past that the convention's inspections could threaten trade secrets in its biotechnology industry. A more sensitive issue may be the US biological defence programme, says Amy Rossi of the Federation of American Scientists in Washington. "We don't like the idea that someone could inspect our national labs."

China and Iran both made objections to the draft text in May, mainly over export restrictions. But as the current series of talks got under way on 23 July, both countries said they hoped the protocol would be finalized this year.

They and others may be waiting for the United States to kill the protocol to save themselves the embarrassment, says bioweapons expert Milton Leitenberg of the University of Maryland in College Park. "The Americans and the Chinese are playing this awful game of who's going to get the blame."

The United States is developing a range of measures to counter bioweapons and seems intent on relying on these defences rather than backing the convention. But Matthew Meselson, a molecular geneticist at Harvard University and an adviser to the US government on chemical and biological weapons issues, warns that such an approach could augment suspicions that the United States has something to hide. "There is a huge cost if we just walk away and say we'll look out for ourselves," he says.

Author of anti-encryption program faces jail

Rex Dalton, San Diego

A Russian graduate student in computer science faces criminal charges for encryption violations this week in the United States, after his arrest at a computer conference.

The arrest on 16 July in Las Vegas, Nevada, of Dmitry Sklyarov, a doctoral student at the Moscow State Technical University, is the latest clash between authorities and computer scientists concerning US laws involving encryption technology and media interests.

Sklyarov is in trouble not for what he said in his talk, but for his role in promoting encryption-busting software. Groups supporting more free access to encryption technology nonetheless condemned the arrest, which was carried out by the Fed-

eral Bureau of Investigation (FBI) at the Defcon-9 conference of computer scientists and hackers.

If convicted by the federal court in San Jose, California, Sklyarov faces a fine of up to \$500,000 or a maximum of five years in prison. He was investigated by a federal 'high-tech squad' of agents based in San Jose. At a meeting with computer-industry executives in Silicon Valley on 20 July, John Ashcroft, the US attorney, said he plans to set up more of these squads throughout the United States.

Sklyarov faces a felony charge of "trafficking in a product designed to circumvent copyright protection measures" in violation of the 1998 Digital Millennium Copyright Act. He is the author of a computer program



called the Advanced eBook Processor that is distributed by a Moscow-based firm, ElcomSoft. ElcomSoft circumvents the encryption of a program — the Adobe Acrobat eBook Reader — sold by San Jose-based Adobe.

Adobe security officials told FBI agents on 26 June that the ElcomSoft program converts their encrypted product into a 'naked file' in standard portable document format (PDF) that could be transferred to anyone, and explained how this would hurt Adobe's business. A criminal complaint was issued on 7 July, with agents arresting Sklyarov after his talk on the "password protection aspects of electronic books and documents... and PDF format".

"His speech was simply a part of his doctoral work," says Alexander Katalov, ElcomSoft's president, who was in California last week seeking Sklyarov's release on bail. Katalov says that another Russian ElcomSoft employee was held briefly in Las Vegas, but was allowed to return to Moscow.

The San Francisco-based Electronic Frontier Foundation (EFF) criticized Sklyarov's arrest and planned a series of protests on 23 July. In a statement, EFF's executive director Shari Steele complained of "shameful and opportunistic actions against an individual who was here simply to share his knowledge and technical expertise with American scientists".

The EFF is also helping Princeton University computer scientist Ed Felten and colleagues with their legal battle to publish encryption techniques involving music files (see *Nature* 411, 5; 2001). Felten has asked the federal court in San Jose for permission to publish an article that the Recording Industry Association of America, a trade group, asserts will damage music distributors.

The Russian consulate in San Francisco says it has been in contact with Sklyarov in jail, and that he was well and was being provided appropriate legal rights. Adobe officials declined to comment.

NASA mission has wind in its sails

William Triplett, Washington

With the launch on 30 July of its Genesis spacecraft, researchers at NASA's Jet Propulsion Laboratory in Pasadena, California, will begin a three-year, \$209-million study of the origin of the Solar System.

A Delta II rocket will take Genesis about a million miles from the Earth, to a position from which it will fly in a series of loops to collect particles from the solar wind. These will be sorted on board

and brought back to Earth for analysis.

The haul, mainly of ions that have drifted away from the Sun's outer atmosphere into space, is expected to contain evidence of the solar nebula that marked the beginning of the Solar System around five billion years ago.

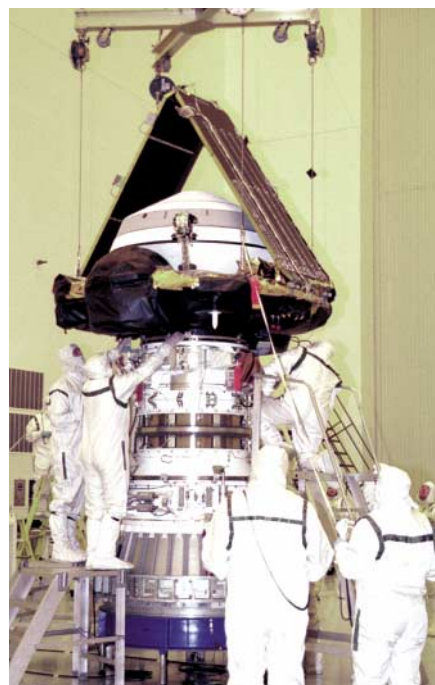
"The outer layer of the Sun preserves a fossil record of the chemical and isotopic composition of the solar nebula, the disk of gas and dust from which all the planets formed," explains Donald Burnett of the California Institute of Technology, the principle investigator on the Genesis mission. The solar nebula is important, he says, because it was the transition phase between when "everything was just stardust" and the matter that now makes up the Solar System. If it succeeds, Genesis will provide insight into what happened during that transition.

John Leibacher of the National Solar Observatory in Tucson, Arizona, chairman of the American Astronomical Society's solar physics division, says that the materials in the solar wind "haven't undergone significant nuclear transformation in five billion years".

Some of the Apollo Moon missions captured solar wind, but on a much smaller scale. Genesis will capture 10–20 micrograms of material over two years, using four instruments: collector arrays, an ion monitor, an electron monitor and an ion concentrator. "We have much purer materials to collect solar wind in now and we have much greater analytical capabilities," says Burnett.

NASA also plans a spectacular homecoming for Genesis. Concerned that a parachute landing could disturb the samples, the agency will deploy helicopters to capture the spacecraft as it floats down. ■

► <http://genesismission.jpl.nasa.gov>



Starter's orders: the Genesis mission aims to unlock the secrets of the Solar System's birth.

Physicists rally behind linear-collider plan

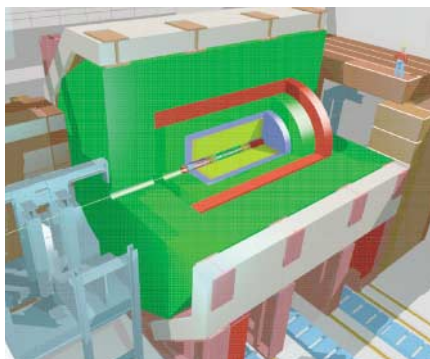
Colin Macilwain, Washington

US high-energy physicists, gathered at a three-week meeting in a ski resort in Colorado, gave their backing to the idea that the construction of a large linear collider is the top priority for their discipline.

"There's widespread agreement that we should be pushing for a linear collider," says Chris Quigg, a theorist at Fermilab in Batavia, Illinois, and co-chair of the conference at Snowmass. The meeting has been organized by the American Physical Society every five years since 1982, and sets the agenda for the discipline in the United States.

Meeting organizers say that even physicists exploring other concepts for large physics machines — such as a muon collider or a very large hadron collider — now see a linear collider as the top priority. Alexander Chao, an accelerator physicist at the Stanford Linear Accelerator Center (SLAC) in California and co-chair of the meeting, says "a majority of people in these two camps would agree" that the linear collider comes next.

But this consensus among US physicists has a downside that was dramatically exposed during the meeting, which ran from 30 June to 20 July. It means that Fermilab is emerging as the most viable US site for the machine, even though Japan has been collaborating with SLAC for several years on the design, on the basis that it might be built in Japan or in California.



First in line: technology developed at DESY in Germany may be used in a new linear collider.

Hiroataka Sugawara, director of KEK, the Japanese high-energy accelerator organization, declared at the meeting that Illinois was an unsuitable site for the collider, and attacked SLAC for appearing to endorse consideration of its location there.

Physicists say Sugawara was angry because he has long suppressed the preference of some KEK scientists for a design that operates in one frequency band, in order to collaborate with SLAC on another.

Chao says there is no doubt that KEK felt betrayed by SLAC. He adds that "the rift between SLAC and KEK strengthens the case" for the technology being developed by Germany's high-energy physics laboratory, DESY (see *Nature* **410**, 397; 2001).

Yukihide Kamiya, KEK's accelerator director, says Sugawara's comments were intended to encourage SLAC to continue pursuing a Californian site for the machine. SLAC's own site cannot accommodate the accelerator. "There is no change in KEK's policy, no change in our desire to collaborate with SLAC in the project," adds Kamiya.

In the past, some US physicists have expressed concern that the linear collider will be too expensive to build, and suggested that other ideas, such as a muon collider, might become their top priority. And many university physicists remain concerned about the impact of a large construction project on their grant funding.

But soundings at Snowmass, including an informal poll of the 1,200 attendees, found strong support from all segments of the community for the linear collider, whether it is built in the United States or elsewhere.

The High Energy Physics Advisory Panel, a group of scientists that advises the US Department of Energy and the National Science Foundation, is preparing a 20-year plan for the discipline, with an interim version of its report due in October.

"The community feels that high-energy physics needs a new facility, and the majority think that it should be a linear collider," says Barry Barish of the California Institute of Technology, co-chair of the advisory panel writing the report.

DESY/MEA

Sea lions massacred in Galapagos for sex organs

Rex Dalton, San Diego

About 25 sea lions in the Galapagos National Park were butchered for their sexual organs earlier this month, angering biologists who work to conserve them.

The mutilated sea lions of both sexes were found on the morning of 14 July on a beach on San Cristobal Island in an area near the airport where tourists go to view the island's diverse animal life. The sea lions had been clubbed in the head the night before, and their penises, gonads or ovaries cut out. Some of the animals were still alive when they were found.

Michael Bliemsrieder, a US-based biologist with the Charles Darwin Foundation, says sea-lion sex organs are sold in Asia for medicinal purposes.

This year's sea-cucumber harvest is under way in the islands' waters, attracting many foreign traders. Bliemsrieder says Ecuadorian authorities are investigating whether any of the traders were involved in the attack. There have been isolated attacks on sea lions during previous harvests,

officials say, but nothing on this scale.

Paola Diaz, a spokeswoman for the Charles Darwin Research Station on the

islands, says scientists and the community have raised around US\$5,000 for a reward for information leading to an arrest.



Mutilated: these sea lions were killed for their sex organs, which could end up on sale in Asia.

CDRS/GNPS



Fruitful work: sequencing the banana genome may lead to disease-resistant varieties.

International body aims to unpeel the banana genome

Rome The genome of the main ancestral species of banana could be sequenced within five years, according to plans announced last week by the International Plant Genetic Resources Institute.

Bananas are one of the developing world's most important crops, but because almost 90% are produced by small farmers and consumed locally, they have benefited little from advances in crop science. The genome data, which will be made available free of charge to non-commercial interests, will help to identify genes involved in disease and pest resistance, and delayed ripening.

The banana is also of interest because it has undergone between 6,000 and 8,000 years of celibacy since the original founder species were bred. Bananas reproduce asexually so, apart from random mutations, the fruit's genetic make-up has remained unchanged. By comparing wild and cultivated strains, geneticists hope to discover how these mutations have helped wild varieties to adapt to pests and diseases.

♦ <http://www.ipgri.cgiar.org>

Top senator pushes new stem-cell compromise

Washington President George W. Bush was offered a possible way out of his stem-cell research imbroglio last week, when an influential Republican senator laid out a set of possible conditions for government funding.

Bill Frist (Republican, Tennessee), the only physician in the Senate, said on 18 July that the research should be federally funded "within a carefully regulated, fully transparent framework [that ensures] the highest level of respect for ... the human embryo".

Frist added that research on embryonic stem cells had greater potential than studies on adult stem cells. But his ten-point plan for

the research would restrict it to an unspecified, finite number of embryonic stem-cell lines — an idea that many biologists oppose (see *Nature* 412, 107; 2001.)

Senate plan favours big-ticket projects

Washington The US National Science Foundation's funding suffered a reverse last week when a Senate committee voted to increase it by 5.6% in 2002. A more generous increase of 9.5% had been approved by a House of Representatives subcommittee panel only a week earlier (see *Nature* 412, 259; 2001).

A few programmes were singled out for large raises, however. A figure of \$296 million was suggested for information-technology projects, up 14% from last year. And the Atacama Large Millimeter Array in Chile would receive \$12.5 million to allow construction to begin — provided that the project will be tightly managed.

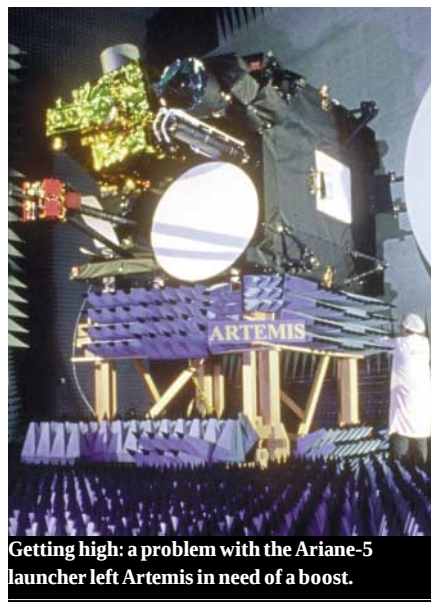
Members of the Senate and the House will meet to reconcile their spending plans before the end of September.

Stray satellite needs help in finding its true calling

Paris The European Space Agency (ESA) has begun attempts to reposition the Artemis telecommunications satellite, which was launched into the wrong orbit on 12 July.

Artemis is the most advanced telecommunications satellite yet developed by ESA. It should have been sent into a geostationary orbit 36,000 km above the Earth, but a problem with the upper stage of the Ariane 5 launcher left the satellite in a much lower orbit.

ESA scientists and engineers from



Getting high: a problem with the Ariane-5 launcher left Artemis in need of a boost.

Alenia Spazio, the Italian company that built the satellite, have already moved the satellite into another orbit much closer to the correct position. They plan to make the final move after allowing three months to check that the payload is functioning properly. This should leave the satellite with enough fuel to function normally — but ESA doesn't know for how long.

Beating surgical vCJD risk may prove costly

London Researchers at Imperial College, London, have identified a possible risk of transmission of the human form of BSE, variant Creutzfeldt-Jakob disease (vCJD), through surgical procedures on the rear of the eye and on the rectum (Wadsworth, J. D. F. *et al.* *Lancet* 358, 171–180; 2001).

By extending an existing prior test, the team were able to test peripheral body tissues for the prions that are thought to cause vCJD. The test confirmed previous studies which showed that the highest concentrations of prions outside the brain are found in the tonsils. But it also showed that low levels are present in the optical nerve, the retina and, in one case, rectal tissue.

Britain is currently introducing a policy of using disposable instruments in tonsil surgery, because of fears that infectious prions cannot be removed from surgical equipment. Operations on the rear of the eye are rare, but the use of disposable instruments in rectal biopsies, which are more common, could prove costly. The Department of Health said it would seek advice on the implications of the new study.

Tortoise passports could save US cattle herd

Washington Faced with a deadly threat to the nation's cattle, the US Department of Agriculture (USDA) had to act. The solution? Passports for tortoises.

Restrictions on the reptiles' right to roam were finalized last week, following fears that rare breeds of tortoise imported from Africa could be host to ticks carrying heartwater, a deadly cattle disease. African livestock have some resistance to the disease, but officials estimate that a US outbreak could kill 80% of its cattle.

Imports of the tortoises have been banned, but USDA officials fear that animals already in the country could spread the disease if their owners moved them around. An initial total ban on tortoise travel brought protests from pet owners, so the USDA now says that it will supply travel documents to tortoises that have been certified as tick-free.

Building a biopolis

Having established itself as a financial and manufacturing centre, Singapore now wants to become a leading player in advanced biological research. David Cyranoski assesses the scientific ambitions of a vibrant city-state.

Clean, confident and cosmopolitan — that is the impression Singapore gives most visitors. With English as its first language, many foreigners quickly feel at home. And under the watchful eye of its autocratic but competent government, Singapore has become one of Asia's leading financial and industrial centres.

Now the ambitious leaders of this city-state have turned their attention to biological science. They plan to make Singapore a major player in biomedicine and biotechnology — and as the cash begins to flow, expectations are rising. Some Singaporean biologists talk excitedly about showing developing countries how to compete with the world's best. Others see their city becoming the hub of a pan-Asian research network. Those with an eye for business, meanwhile, are promoting it as a gateway to the lucrative Asian market for Western drug and biotech companies.

This push in biology is making waves internationally. "They have really bright, far-sighted people in high places willing to put in significant investment," says pharmacologist Paul Lietman, research director of the Singapore branch of Baltimore's Johns Hopkins University. "Singapore could and should be a model for other nations that are trying to build biomedical research capacity."

Rising tension

But if Singapore's biological ambitions are to be realized, some nagging problems must be addressed. Foremost are concerns about the recruitment and retention of sufficient top-class researchers. "The rate-limiting step is getting the people," says Sydney Brenner of the Salk Institute for Biological Studies in La Jolla, California, one of a group of prominent biologists advising the Singaporean government. And in the current hothouse atmosphere, rivalries between disciplines and institutes have created some bitter tension.

The driver behind Singapore's scientific revolution is Philip Yeo, who took over as chair of the National Science and Technology Board (NSTB) in February. Having headed the Singapore Economic Development Board since 1986, Yeo gained credit for the rise of the city's microelectronics and other lucrative

industries — giving him a celebrity status unusual for a government official. Nowadays, his efforts are concentrated on the expansion of Singaporean biomedicine. Trained as an industrial engineer, Yeo brings in biologists to tutor him on Saturday afternoons. When *Nature* visited his office, the previous week's notes — on cell signalling — were still scrawled on a whiteboard.

"We're building the entire range of medical advance, from drug discovery through early-stage clinical trials to manufacturing and marketing," says Yeo. The government is putting money behind this vision. The NSTB's budget from 2001 to 2005 is some US\$3.8 billion, a 75% increase over the previous five years. Most disciplines are benefiting, but biomedicine has been singled out. The Biomedical Research Council (BMRC), established last September, has a five-year budget of some US\$800 million, and is ploughing these funds into a series of new institutes.

The NSTB is also building a chemical institute as infrastructure for its budding pharmaceutical and biotechnology industries. Major biotech players such as Chiron of Emeryville, California, have announced plans to set up research initiatives in Singapore. Last month, the drugs multinational Eli Lilly and the Singapore Economic Development Board announced the opening of a joint Center for Systems Biology, with a research budget of US\$140 million over five years.

In this favourable climate, Singaporean biology is making rapid progress. The



Mapped out: the city's leaders plan a glittering biological future.

nation's small size means that there is little inertia. "We can turn on a dime,"

boasts Chris Tan, director of the Institute of Molecular and Cell Biology (IMCB), which is a key player in the international effort to sequence the genome of the puffer fish *Fugu rubripes*.

The IMCB is now part of the BMRC's expanding empire, as is the Genomics Institute of Singapore, created with an influx of funds announced a year ago. Its executive director is Hong Kong-born Edison Liu, former head of clinical sciences at the US National Cancer Institute in Bethesda, Maryland. With an operating



Leading lights: Chris Tan, inset, and his Institute of Molecular and Cell Biology.

budget of US\$25 million a year, Liu plans to recruit 250 researchers to study areas such as the transcription of genes into proteins and the influence of genetic variability on drug metabolism within Singapore's population.

To manage the anticipated deluge of genomic data, the BMRC in February added a new Bioinformatics Institute — expanded from the National University of Singapore's existing bioinformatics centre. And the BMRC is now gearing up for the launch of a bioengineering centre to work on tissue engineering and regenerative medicine.

At present, the BMRC's institutes are hosted by the National University. But in 2003, they are scheduled to move into a science park dubbed the 'Biopolis', being built in the district of Buona Vista, still close enough to interact with the university. The Biopolis will include living space for up to 1,000 scientists. "It will be a place where people want to come," says Louis Lim, the BMRC's executive director, who has his sights set on attracting the best international talent.

Thin veneer

For the time being, that will be essential, as there are not enough home-grown scientists to fill all the available positions: more than three-quarters of the researchers at public institutes are foreigners. The English language — a legacy of British colonialism and the glue that holds together Singapore's Chinese, Malay and Indian communities — is an important draw. For researchers from elsewhere in Asia, Singapore provides an opportunity to become comfortable with English while still in a familiar culture. For Westerners, the language makes integration easier.

But some observers say that too many visiting researchers view Singapore as a temporary stopover, which means that the city-state gains less than it should from its investment in science. "Singapore's research infrastructure is a thin veneer," one distributor of reagents and scientific instruments told *Nature*. "There is a lot of top-of-the-line equipment, but few know how to use it."

Yeo and other leaders recognize the problem, and are taking steps to solve it. The Bioinformatics Institute, for instance, plans to churn out 100 graduates each year. The National University of Singapore and the Nanyang Technological University, meanwhile, have established a collaboration that will see 40 leading professors at the Massachusetts Institute of Technology offer masters courses over the Internet for 30 Singaporean students each year.

With these and various other programmes — for attracting foreign researchers, training Singaporeans and bringing back natives who have trained abroad — Lim is optimistic about the future. "Five or ten years down the line, there will be no shortage of talent," he claims. The IMCB's Tan argues that Singapore should not worry too much about

researchers leaving for foreign labs, provided they stay long enough to make a significant contribution. "Ideally, we'd be a 15-year revolving door," he says.

Some foreign scientists find it hard to adjust to a country sometimes dubbed "Singapore Inc.", in which the government exerts control over most aspects of life — in particular those that underpin the nation's economic competitiveness. "It's the most hierarchical place in Asia," claims one visiting biomedical researcher. But other scientists argue that Singapore no longer deserves its reputation for control-freakery. "The government does not direct my research at all," says Liu. "I can exercise my own vision."

For a visiting journalist, perhaps the biggest surprise is the bitter rivalry between some of Singapore's leading institutes. Tan's IMCB stood alone for eight years after its launch in 1987 as the city's world-class biology centre. But in 1995, it gained some competition when plant molecular biologist Nam-Hai Chua of Rockefeller University in New York convinced the NSTB to establish the Institute of Molecular Agrobiology (IMA).

The IMA now has 190 researchers and an enviable scientific reputation. But some powerful critics say a tiny, densely populated island has no place investing in agricultural research. "It's a criminal waste of taxpayers' money," says Yeo. "I would close it down tomorrow if I could." To stress his point, Yeo suggests a trip to a supermarket. "Everything there is imported. We have no agriculture."

Such violent attacks have placed the IMA's soft-spoken director, Venkatesan Sundaresan, in a difficult position. Last autumn, it looked as if the IMA's days might be numbered, as articles in *The Straits Times*, Singapore's leading newspaper, debated the wisdom of investing in agricultural biotechnology. But in November, the outlook brightened when the IMA gained a powerful



Precious resource: Singapore's biggest problem in biology is recruiting and retaining staff.



Only the best will do: Singapore's laboratories are furnished with top-line equipment.

advocate in its newly appointed chair, Ho Ching, chief executive of the engineering conglomerate Singapore Technologies.

Whether the rivalry will harm Singapore's efforts to compete internationally is unclear. But many researchers would prefer the protagonists to collaborate in pushing for Singapore to become the centre of an emerging research network. Established in 1997 by a group of leading Asian biologists, the Asia-Pacific International Molecular Biology Network (IMBN) is envisaged as a counterpart to the successful European Molecular Biology Organization.

So far, the IMBN has organized a handful of meetings and assembled a membership of almost 250 scientists. But its leaders are now pressing their respective governments to provide the funds to run a fellowship programme and establish a laboratory — probably in Singapore, which already hosts the organization's secretariat. "Singapore is the natural home for the IMBN," says Ken-ichi Arai, president of its governing council and director of the University of Tokyo's Institute of Medical Science.

Arai and his colleagues face a tough task winning financial backing for their vision (see Opinion, page 361). But if they do attain their goal of establishing an Asian analogue of the European Molecular Biology Laboratory in Singapore, it would set the seal on the city-state's emergence as a biological power. ■

David Cyranoski is *Nature's* Asian-Pacific correspondent.

National Science and Technology Board

► <http://www.nstb.gov.sg>

Institute of Molecular and Cell Biology

► <http://www.imcb.nus.edu.sg>

Genome Institute of Singapore

► <http://www.genomeinstitute.org>

Institute of Molecular Agrobiology

► <http://www.ima.org.sg>

Asia-Pacific International Molecular Biology Network

► <http://www.a-imbn.org>

IMA

NSTB

Picture perfect

Medical imaging techniques are being adapted to study gene expression and other cellular activities in living animals. Corie Lok talks to the pioneers who are watching cells at work in their natural habitat.

FROM REF. 2

A decade ago, Harvey Herschman was a classic laboratory gene-hunter, focused on studies of cell cultures. His team at the University of California, Los Angeles (UCLA), was internationally respected. Yet Herschman felt he was missing something important.

"There's always something unsatisfactory about studying genes *in vitro*," he says. The same is also true for studies on laboratory animals, he adds, if all you can gain is a frozen snapshot of gene activity by doing post-mortems or biopsy studies. So in 1996, Herschman teamed up with imaging specialists, including his UCLA colleague Michael Phelps, to devise techniques to watch gene expression in living animals.

Herschman is one of a growing number of leading cell biologists, molecular biologists and geneticists who are working with experts in magnetic resonance imaging (MRI), positron emission tomography (PET) or optical imaging. "Four years ago, I couldn't spell PET," jokes Herschman. Today, he is principal investigator at UCLA's Center for *In Vivo* Imaging in Cancer Biology.

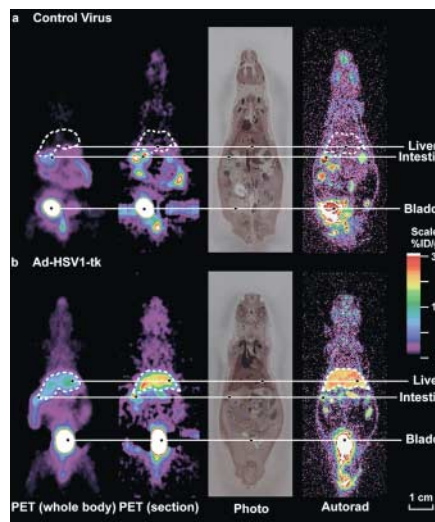
Glowing report

Advances in molecular and cellular biology have yielded a host of molecular targets, including genes and proteins, that researchers would like to study in living animals. The field of *in vivo* molecular imaging is being driven by the development of chemical probes to make these targets 'light up' in MRI, PET or optical images, and by improvements in the techniques' resolution.

The resulting images promise new insights into areas such as embryonic development, infectious disease and gene therapy. They should also allow researchers to study the subtle influences that environmental factors exert on gene expression. "Imaging brings the genome to life," says Phelps, who helped pioneer the development of PET in the 1970s. "It's powerful because it puts all the pieces together."

That power has been demonstrated in the past few years by proof-of-concept papers showing that MRI, PET and optical techniques can non-invasively visualize where, when and, in some cases, at what level a gene is being expressed in living animals.

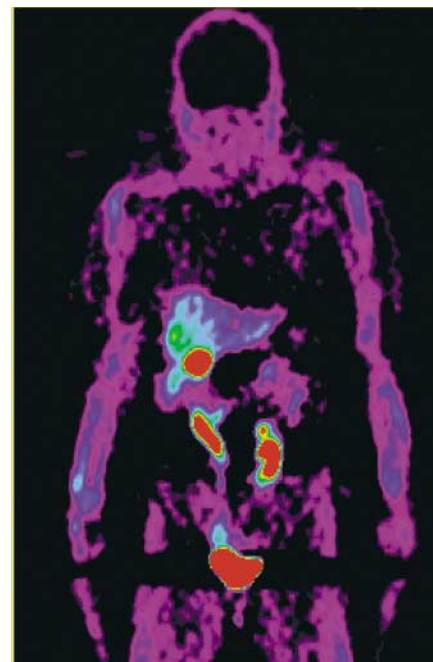
Juri Tjuvajev, Ron Blasberg and their colleagues at the Memorial Sloan-Kettering



Colour-coded: positron emission tomography can be adapted to reveal the levels at which introduced genes are being expressed in mice (above) and human patients (right).

Cancer Center in New York began exploring methods for imaging gene expression almost eight years ago. One goal was to monitor the effectiveness of gene therapy. Determining whether the genes introduced during therapy are active in the target tissues is key to judging the treatment's success, and techniques that rely on taking multiple biopsies are clearly undesirable.

Tjuvajev and Blasberg turned to PET, which produces images by detecting the radiation given off by molecules 'tagged' with radioactive isotopes. In 1998, their team published the first report using PET to



monitor gene expression in rats¹. The gene that they studied, *HSV1-tk*, comes from the herpes simplex virus 1. It produces an enzyme that converts the drug gancyclovir from an inactive form into a form that kills cells. In what is dubbed 'suicide gene therapy', researchers hope to combat cancer by selectively introducing the gene into tumour cells and then administering the drug.

The enzyme produced by *HSV1-tk* adds phosphate groups to thymidine—one of the four chemical 'bases' in the genetic code—and related molecules. The addition of phosphate to a thymidine analogue called FIAU



Clear-sighted: Harvey Herschman (left) and Sanjiv Gambhir have teamed up with Michael Phelps (opposite) to develop PET technology to visualize gene activity in laboratory animals.

S. YAGHOUBI ET AL. / J. NUCL. MED. (IN THE PRESS)

UCLA'S JONSSON CANCER CENTER

converts it into a form that gets trapped inside cells. So by injecting their rats with radioactively labelled FIAU, Tjuvajev and Blasberg were able to use PET to show where, and at what level, the introduced *HSV1-tk* gene was being expressed.

A year later, a team in Herschman's lab led by Sanjiv Gambhir repeated Tjuvajev and Blasberg's success in mice². Rather than using a standard clinical PET machine, Gambhir and his colleagues took advantage of a scanner with higher resolution designed by UCLA colleagues led by Simon Cherry specifically for imaging small animals³. This device, called the microPET, is now being manufactured by Concorde Microsystems, a company in Knoxville, Tennessee.

Into the clinic

PET imaging of *HSV1-tk* expression is now being used in a clinical gene-therapy trial. Wolf-Dieter Heiss and Andreas Jacobs of the Max Planck Institute for Neurological Research in Cologne and the University of Cologne have administered the *HSV1-tk* gene, packaged inside tiny balls of lipid, into recurring malignant brain tumours in six patients. They are using PET to monitor the gene's incorporation into tumour cells.

The Sloan-Kettering and UCLA groups have also used the *HSV1-tk* system to monitor the expression of a second added gene, which should allow the gene to be used as a general marker for the success of gene therapy. By linking both genes to the same regulatory DNA sequence, they get expressed in unison. In 1999, Tjuvajev and Blasberg demonstrated the principle⁴ using single-photon emission computed tomography, a technique related to PET that uses different types of radioactive probes. The following year, the UCLA researchers reported similar success using PET⁵.

More recently, both groups have adapted the system to study the expression of the body's own 'endogenous' genes, rather than those added in gene-therapy experiments. The activity of a gene can be studied indirectly by taking the regulatory sequence of DNA that controls its activity and linking it to a

marker gene that produces an easily detected signal. By creating transgenic mice in which *HSV1-tk* is linked to the regulatory sequence for the gene encoding the protein albumin, the UCLA team has shown that the system can monitor endogenous gene expression⁶. The Sloan-Kettering researchers, meanwhile, have created mice bearing tumours in which *HSV1-tk* is linked to a regulatory sequence that is controlled by the p53 protein, which suppresses tumour growth. Because p53 turns on a series of other genes, the resulting images can be used to monitor their expression⁷.

MRI, which has a higher resolution than PET, can also be adapted to monitor the behaviour of genes and proteins. The technique measures changes in the magnetization of hydrogen protons in water molecules sitting in a magnetic field after they have been hit with a pulse of radio frequencies. Protons from different tissues react differently, giving a picture of anatomical structures. These images can be enhanced by adding 'contrast agents', which sharpen the contrast by affecting the behaviour of protons in their proximity. In standard clinical MRI scans, these agents travel through the bloodstream and tissues, heightening contrast wherever they go. To image cellular activities, researchers are tailoring the agents so that they are only 'turned on' in the presence of a specific molecule.

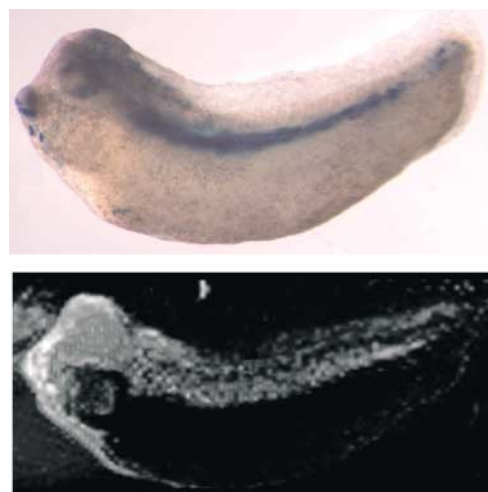
Two teams, one led by Ralph Weissleder at the Massachusetts General Hospital in Boston, the other by Thomas Meade of the California Institute of Technology in Pasadena, are making rapid progress.

Contrast control

Weissleder and his colleagues have spotted the expression of a receptor gene engineered into tumour cells that were then implanted into mice. They took nanoparticles of iron oxide and chemically linked them with the protein that binds to the receptor so that only clumps of cells with the receptor on their surfaces lit up in the resulting images⁸. Contrast agents could be targeted to other cell-surface proteins in the same way, Weissleder suggests.

Meade's team lit up gene activity in frog embryos by altering another MRI contrast agent — a gadolinium ion inside a chemical scaffold — so that it was activated only by the product of the gene in question, an enzyme called β -galactosidase. By attaching sugar molecules to the contrast agent, the researchers switched off the molecule's ability to interact with nearby protons. But the β -galactosidase enzyme removed the sugar groups, turning up the contrast wherever the gene was active.

To demonstrate the technique, the researchers injected the gene for β -galactosidase, known as *lacZ*, into one of two cells in newly fertilized embryos of *Xenopus laevis*,



Magic marker: in these frog embryos, *lacZ* gives a blue colour (top) and a clear MRI signal.

the African clawed frog, that had undergone one cell division. Both cells were injected with the contrast agent. As the embryo grew, the side of the animal that had developed from the *lacZ*-injected cell glowed in the MRI images⁹. Scott Fraser, a developmental biologist on the team, believes that MRI will allow researchers to watch cells move, divide and specialize as an embryo grows. "Seeing where and when molecular events take place is critical in developing systems," he says.

The *lacZ* gene is often added as a 'marker' together with the gene of interest in transgenic experiments, as β -galactosidase produces a blue precipitate when it acts on a molecule called X-gal. Meade's technique should allow its role as a marker to extend to watching the incorporation of genes deep within a living animal's tissues. More generally, the Caltech team is working to develop and refine contrast agents that are activated by other enzymes, such as those produced by tumours, and by calcium ions, which act as messengers in many important cell-signalling pathways¹⁰.

MRI and PET require expensive equipment that is beyond the reach of many labs. But in some cases, it is possible to track gene activity in lab animals using optical tech-



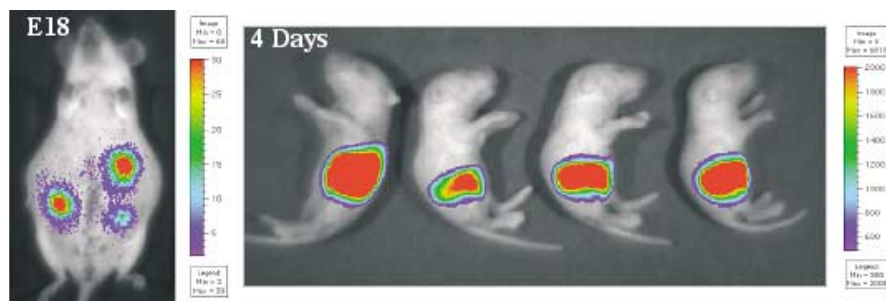
Pioneer: Ron Blasberg's team produced the first PET images of gene expression in live rats.

UCLA



FROM REF. 9

MSKCC



In the family way: optical imaging can monitor gene expression in mice before (left) and after birth.

niques. One approach depends on genes for enzymes called luciferases, found in fireflies and certain bacteria, which generate light by oxidizing a molecule called luciferin. If this glow is sufficiently bright, light produced deep within an animal following the administration of luciferin can be detected by sensitive charged-couple-device (CCD) cameras.

Caught on camera

Researchers at Stanford University in California, led by Christopher Contag, his wife Pamela and David Benaron, have pioneered this approach, having first turned to optical imaging while trying to monitor the movement of bacteria and viruses in animals. They were frustrated with traditional techniques of analysing biopsies for the presence of pathogens, which were slow and cumbersome. "We had questions and the only way to answer them was to build new technology," says Christopher Contag.

The researchers built a prototype CCD-based camera system in 1994 and began using it to study mice infected with bacterial pathogens that had been engineered to carry the luciferase gene¹¹. In 1997, Pamela Contag left Stanford to form Xenogen, a company that has further developed the technology.

The Stanford team has shown that the system could also be used to monitor the expression of endogenous genes indirectly. Using transgenic mice carrying a luciferase gene linked to a regulatory sequence that is activated by the chemical dimethyl sulphide, the researchers confirmed that luciferase can be used as a marker to study gene expression deep within a mouse's body¹². Having proved the principle, it should be possible to study endogenous gene expression using marker-gene strategies similar to those used by groups working with PET.

The Stanford researchers have also joined forces with Karin Gaensler of the University of California, San Francisco, to use luciferase in tests of a viral vector being developed for gene therapy on fetuses. Gaensler put a luciferase gene into the virus, and injected it into mouse fetuses *in utero*. For months after the pups were born, CCD cameras detected luciferase activity in their tissues¹³.

In Boston, meanwhile, Weissleder has developed another optical-imaging technol-

ogy to study tumour biology. Near-infrared fluorescence uses longer wavelengths, as these can penetrate through greater depths of tissue. Weissleder and his team have designed molecular probes that fluoresce at near-infrared wavelengths only in the presence of specific protein-digesting enzymes produced at high levels by tumours. These enzymes are thought to be involved in tumour growth and spread, plus the formation of the blood vessels that supply growing tumours with nutrients and oxygen.

The researchers linked fragments of proteins that would be digested by the tumour enzymes to the fluorescent molecules in a way that turned off their near-infrared glow. They then injected these probes into mice suffering from cancer. The tumour enzymes got to work on the protein fragments, releasing the fluorescent molecules and lighting up the tumours¹⁴.

More recently, Weissleder's team injected a probe activated by matrix metalloproteinase-2 (MMP-2), a key protein-digesting tumour enzyme, into mice with tumours and showed that the near-infrared fluorescence signal decreased after an inhibitor of the enzyme was administered¹⁵. MMP-2 inhibitors are being investigated as cancer drugs, and Weissleder suggests that near-infrared imaging could provide feedback about their efficacy.

Optical techniques have so far been limited to producing two-dimensional images, rather than the three-dimensional reconstructions that are possible with MRI or PET. But Weissleder is now working on computer

techniques to generate three-dimensional near-infrared images¹⁶.

As molecular imaging techniques continue to advance, other researchers are becoming excited about the field's potential. Stem-cell biologists, for instance, are intrigued by the possibility of watching stem cells migrate to repair damaged tissues, whereas immunologists are tagging immune cells and watching them move inside lab animals.

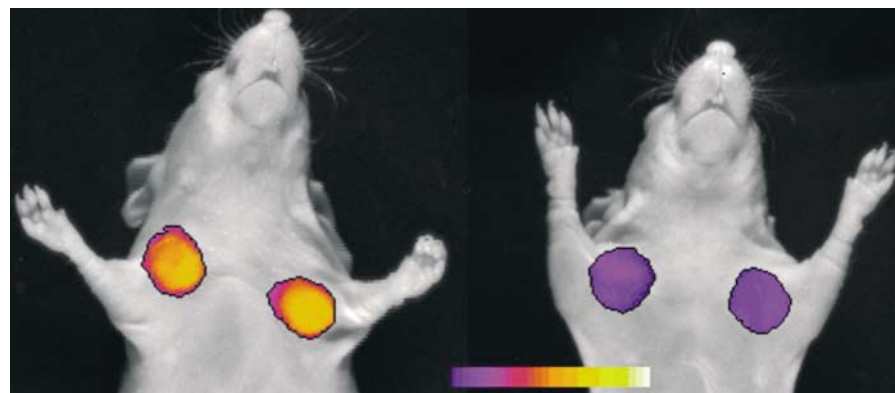
Probing ahead

One major factor determining the field's progress is the development of new probes to home in on particular molecular targets. An important goal is to design probes that home in on the messenger RNA and proteins produced by particular genes, to allow the expression of endogenous genes to be studied directly. Other probes in the works include luciferases that give off different wavelengths of light, which could be used in combination to study the simultaneous expression of different genes; MRI contrast agents that can readily penetrate cell membranes; and PET probes that target specific tumour proteins.

As new probes emerge, enthusiasts such as Herschman are keen to get them into the hands of as many researchers as possible. He hopes to open the doors to his imaging centre to other scientists at UCLA and elsewhere. "They can start thinking about doing experiments in animals that they couldn't do before," he says.

Corie Lok has just completed an internship with Nature.

1. Tjuvajev, J. G. *et al. Cancer Res.* 58, 4333–4341 (1998).
2. Gambhir, S. S. *et al. Proc. Natl Acad. Sci. USA* 96, 2333–2338 (1999).
3. Cherry, S. R. *et al. IEEE Trans. Nucl. Sci.* 44, 1161–1166 (1997).
4. Tjuvajev, J. G. *et al. Neoplasia* 1, 315–320 (1999).
5. Yu, Y. *et al. Nature Med.* 6, 933–937 (2000).
6. Green, L. A. *et al. Mol. Imaging Biol.* (in the press).
7. Doubrovin, M. *et al. Proc. Natl Acad. Sci. USA* (in the press).
8. Weissleder, R. *et al. Nature Med.* 6, 351–354 (2000).
9. Louie, A. Y. *et al. Nature Biotech.* 18, 321–325 (2000).
10. Li, W.-H., Fraser, S. E. & Meade, T. J. *J. Am. Chem. Soc.* 121, 1413–1414 (1999).
11. Contag, C. H. *et al. Mol. Microbiol.* 18, 593–603 (1995).
12. Contag, C. H. *et al. Photochem. Photobiol.* 66, 523–531 (1997).
13. Lipshutz, G. S. *et al. Mol. Ther.* 3, 284–292 (2001).
14. Weissleder, R. *et al. Nature Biotech.* 17, 375–378 (1999).
15. Bremer, C. *et al. Nature Med.* 7, 743–748 (2001).
16. Ntziadristos, U. & Weissleder, R. *Optics Lett.* 26, 893–895 (2001).



The drugs do work: near-infrared fluorescence shows the effects of a cancer drug (treated, right).

Framework welcome, but could do with fine-tuning

Let's get details sorted out, timescales defined, postdocs funded, entrepreneurs trained.

Sir—The European Commission proposal for the sixth Framework programme (FP6) needs to be adopted before December to avoid a break in European research programmes. Several key elements of the proposal deserve strong support, such as networking national programmes; increased coherence of national science and technology policies; increased mobility for researchers; and the long-requested simplification of project management via a decentralized system. And of course, we can only applaud the proposed significant increase in the overall budget for European science and technology research.

Nevertheless, as stated in the Opinion article "Storm clouds over Brussels" (*Nature* **411**, 871; 2001), important aspects of the proposal need to be clarified, and the details of how the objectives are to be put into action have not been properly addressed.

Previous Framework programmes have been beneficial, but have had recurring, still unsolved, administrative problems. Although the FP6 proposals should provide better flexibility and efficiency, it remains to be explained how the projects will actually be managed. Clear timescales for reviewing applications and delivering money should be defined at the outset.

The current Framework system does not provide funding mechanisms or coordinating bodies for Europe-wide basic and strategic science. Your Opinion article suggested as one candidate the coal and steel research fund. The European Science Foundation's Eurocores programme is another, in which case FP6 should contribute to its funding.

Although FP6 provides strong support for the mobility of researchers in the European Union, several categories of scientists remain unable to benefit from this money, including doctoral students. Postgraduate students should be allowed to benefit from this money as they are the key to better collaboration between universities and greater compatibility between doctoral programmes in different European countries. Longer-term funding to help postdocs find their first tenured position should also be made available, as proposed by the Marie Curie Fellowship Association. In addition, the provision of return grants should be developed, to help young European scientists (especially those from Eastern Europe) return to their home countries with decent support.

Feasibility studies, as well as the initial investment and, in some cases, operating

costs of novel research infrastructures, should be funded by the Framework programme. A high-level coordinating body is necessary to ensure the continuity and coherence of policy decision in this domain. International collaboration also appears to be poorly provided for; collaborations with Mediterranean and developing countries in particular are greatly underfunded.

Finally, European start-up companies suffer from the insufficient training of

would-be entrepreneurs in capital fundraising compared with their US counterparts. Training in venture-capital rules should be co-funded by FP6 to help European business innovation. In addition, a European insurance body should be set up to cover the costs of patenting and of defending patents at the international level.

Frédéric Sgard

Euroscience, Sanofi-Synthelabo, 10 Rue des Carrières, 92500 Rueil-Malmaison, France

Climate-change strategy needs to be robust

Sir—Stephen Schneider argues in his Commentary "What is 'dangerous' climate change?" (*Nature* **411**, 17–19; 2001) that the Intergovernmental Panel on Climate Change should assign subjective probabilities to the "storyline" scenarios of future climate change offered in its recent special report on emissions scenarios (SRES), because probabilities will be required by policy-makers.

Schneider raises an important concern, in that a set of plausible scenarios does not, by itself, tell policy-makers what is likely to happen or provide clear guidance for action. But his proposal does not take into account that decision-makers must form a climate policy acceptable to groups with many different, yet plausible, estimates of the likelihood of alternative futures.

This condition of deep uncertainty differs from many risk-management problems, in that little solid information exists to inform subjective probabilities for the long-term economic, social and technological trends underlying different greenhouse-gas emission scenarios. It is unlikely that scientific evidence will soon resolve the assumptions about the socio-economic future made by different groups.

Under such conditions of deep uncertainty, decision-makers often rely on robustness—that is, strategies that work reasonably well compared with the alternatives across a wide range of plausible scenarios (R. J. Lempert & M. E. Schlesinger, *Climatic Change* **45**, 387–401; 2000). They then form a policy solution that works reasonably well for all possibilities. In contrast, an optimum strategy requires probabilities for alternative scenarios, which can invite discord, as well as leading to policies that

can fail in some plausible futures.

In our view, the SRES team was correct to avoid assigning probabilities and instead recommend use of ensembles of multiple scenarios to capture what is known about the long-term climate future. Researchers can now use such scenarios to assess the robustness of alternative, near-term climate policies. Such studies are likely to suggest near-term policy choices that are largely insensitive to many uncertainties. Only when such studies also reveal a few key assumptions to which near-term choices are particularly sensitive, can scientists take the perilous step of putting limits on the range of subjective probabilities for those few assumptions.

Robert Lempert*, Michael E. Schlesinger†

**RAND, 1700 Main Street, Santa Monica, California 90407-2138, USA*

†Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, 105 S. Gregory Street, Urbana, Illinois 61801-3070, USA

When three's not a crowd

Sir—You are showing your zoocentric side with your description on the In This Issue page of the rarity of three-way symbioses ("Ménage à trois", *Nature* **411**, xi; 2001). This sort of arrangement is not so rare among plants: for example, leguminous plants with mycorrhizal associations—see *Lichen Biology* (ed. T. Nash III, Cambridge Univ. Press, 1996). An even more intimate relationship exists among a number of tripartite lichens, where a composite organism is formed through the symbiosis of an ascomycete, a chlorophytic alga and a cyanobacterium. This association can be so specific that, from a practical standpoint, at least two of the individual partners may not exist outside the symbiotic relationship.

Michael A. Thomas

Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9015, New Zealand

Medicine as performance

Early British modern medicine can be considered as a type of theatre.

Bodies Politic: Disease, Death and Doctors in Britain, 1650–1900

by Roy Porter

Reaktion Books/Cornell University Press:

2001. 304 pp. £25/\$35

John Harley Warner

Coarse, vulgar, showy, revelling in crude stereotypes and offensively scatological — such are the images of sick and well bodies and learned physicians and quacks from early modern England that are the core subject matter of *Bodies Politic*. What do they tell us about the culture that produced them? And how can interpreting such images in their historical context transfigure our understanding of the early modern medical enterprise? In Roy Porter's book we have both a free-wheeling romp (with a keen eye for the absurd) through more than two centuries and a cogent scholarly rethinking of post-Restoration bodily beliefs and medical practices.

An engaging confession opens the book. Porter, the doyen of the social history of British medicine during the “long eighteenth century”, confides that it only recently dawned on him that he had never seriously and systematically grappled with visual images. Using pictures as evidence is the explicit aim of the “Picturing History” series in which this volume appears, and 137 illustrations — 39 of them in colour — take pride of place in Porter's analysis. Image-making flourished during the Georgian century, and in this study, etchings, watercolours and lithographs by William Hogarth, James Gillray, Thomas Rowlandson and George Cruikshank occupy centre stage.

Yet the wider focus is not pictures but public representations, and Porter finds his images in verbal as well as visual languages. Aphorisms, novels, proverbs and plays all provide grist to his mill. Using this rich farrago, he shows us the models of practitioners and patients, diseases and treatments, prominently placed on view in the public media. Representations of the body — the grotesque and monstrous body, the healthy and beautiful body, the ailing and doctored body, the dissected body — are vividly displayed and astutely interpreted.

One point of lavishing so much attention on this ‘body language’ is Porter's contention that it was in no small measure through the metaphorical and symbolic meanings attached to the body that early modern Britons made sense of their world and came to terms with its social, political and moral ills. The political arena, as he reminds us, was, and is, rife with corporal allusions and

images so commonplace that we have lost sight of their rootedness in the human frame: the monarch as head of state; Members of Parliament; factional purges and parliamentary motions; and talk of the Constitution. But eighteenth-century cartoonists missed none of the satirical potential, and widespread distrust of doctors made it easy to encode politics through scenes of medical practice. In the hands of Cruikshank, Gillray and Rowlandson, medical encounters depicting patients as hapless yokels subjected to blood-letting, violent purging or surgical dismemberment — scenes that indulged in unsparing caricatures of class and gender — offered mordant social and political commentary.

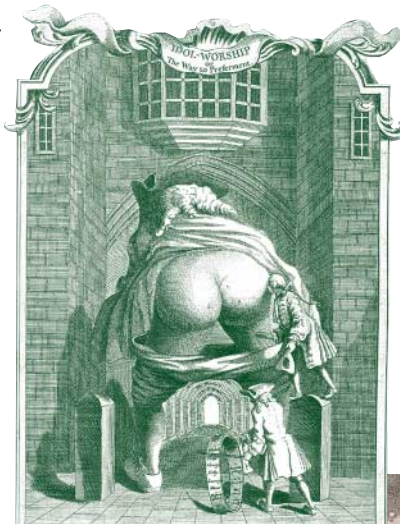
Porter's recognition and explication of the quintessentially performative nature of medicine is perhaps his most deft interpretive stroke. He follows the preoccupation with performance that in recent years has taken cultural studies by storm, but he retains a sturdy grasp on historicity. Porter acknowledges the persisting prominence of medicine in mass entertainment today (he cites television programmes such as *ER* and *Chicago Hope*), the endurance of terms such as operating ‘theatre’, and the more recent sociological construct of the ‘sick role’ consciously or unconsciously adopted by the patient. But he insists that early modern medicine was singularly infused with an awareness, which sufferers and healers shared, that medicine was a type of theatre, of spectacle — an art form in which the practitioner, through rhetoric and ritual, was expected to interpret and give meaning to experience in ways that transcended expectations of cure.

Such a perspective neither dismisses the very real sufferings of the sick nor denies the fact that most patients turned to healers in the hope of a cure. But emphasis on performance forces the historian to pay close attention to the audience, and particularly to the market-place in which medical identities were tried out, refashioned and consolidated. At a time of occupational instability, and with a commercial society in flux, calculated self-promotion, often played out on a highly

visible public stage, was a crucial ingredient in the success of learned physician and mountebank alike. And, indeed, in a Protestant nation in an age of reason, the widespread censure of theatricality — of superficial artifice and mere show — may be an important clue to why public depictions of doctors during the Georgian era were overwhelmingly negative. Although some readers might fault Porter for giving little attention to how technical and conceptual changes within medicine transfigured its public face, his exploration of “the theatre of physic” is crucial to our understanding of early modern healers and healing.

A brief closing foray into Victorian representations of medicine raises tantalizing questions about change over time. Public

Body language: the satirists of the day used the widespread distrust of doctors to offer telling social and political comment.



depictions of medical practitioners, Porter suggests, grew increasingly flattering, just as newer platforms for graphic humour — such as *Punch* (founded in 1841) and *Vanity Fair* (1868) — tended to eschew the coarseness and vulgarity of Georgian and Regency caricature.

Broader cultural shifts plainly fostered these new aesthetic conventions, but the role of changes in medicine itself is less clear. And in Porter's treatment of the nineteenth century, as throughout the book, one would have welcomed something about private representation as a counterpoint to public images and media constructs. If public representations of Victorian healers and healing shed much of their earthiness and theatricality, then what are we to make of certain more covert genres of medical self-portraiture, such as grizzly photographs depicting medical students at work in the dissecting room, or of late-Victorian advice books that unabashedly instructed doctors on how to succeed in business by putting on a good show? If the reader is left wanting more, however, it is because Porter's book is one of the best studies we have of public representations of medicine, a highly visual culture in which spectacle and ritual performance, although transformed, have hardly vanished. ■

John Harley Warner is in the Section of the History of Medicine, Yale University School of Medicine, PO Box 208015, New Haven, Connecticut 06520-8025, USA.

The director's tale

The Recombinant DNA Controversy: A Memoir. Science, Politics and the Public Interest 1974–1981

by Donald S. Fredrickson

American Society of Microbiology Press: 2001. 408 pp. \$39.95, £32.50 (pbk)

Susan Wright

Some 25 years ago, unease within the molecular-biology community over the implications of the emerging techniques of genetic engineering triggered a major public controversy over possible hazards and misuse of the new field. In the United States, the National Institutes of Health (NIH), the world's leading biomedical research institution, was designated as the government agency responsible for developing voluntary controls. The controversy quickly encompassed the nature of the controls themselves, which had been developed largely by scientists whose career (and in some cases business) interests were closely linked to the new field.

The still-dominant view of these events is



View from the inside: Fredrickson (right), seen here with NIH committee member Charles McCarthy.

that the American public, along with a few dissident scientists, overreacted to early concerns about the hazards of genetic engineering. Eventually, the hazards were shown to be exaggerated, allowing the NIH controls to be dismantled. This conventional wisdom casts the story as a confrontation in which rational belief in science ultimately prevailed over irrational fear of the unknown.

Donald Fredrickson, the NIH director who presided over the rise and fall of the NIH controls, elaborates on this story with his legendary panache. He covers the struggles of the NIH not only with the attentive public but also behind the scenes, with the executive and legislative branches of the US government. This previously invisible face of the recombinant DNA controversy is now, apparently, revealed in Fredrickson's papers and personal letters and diaries being deposited in the US National Library of Medicine. This collection promises to substantially augment the public record provided by the NIH and the important Recombinant DNA History Collection at the Massachusetts Institute of Technology archives.

The memoir leaves little doubt about Fredrickson's own view of the genetic-engineering question as he describes the adroit manoeuvres of the NIH to protect "the gossamer quality of [its] pseudo-regulations" from conversion by the US Congress into mandatory controls that encompassed the private sector and all government agencies. Battle metaphors permeate this account of the struggle between those deemed to be 'for' or 'against' the voluntary policy espoused by the NIH. As he wrote in an unpublished article on the eve of the first major weakening of the NIH controls: "I think we have won a significant victory over a dangerously excessive reaction first set in

motion by scientists." The fact that policy-making in certain other countries — notably the United Kingdom — took the turn towards uniform regulation without a bitter struggle is not noted here. This is very much an American story, told by the person at the epicentre of the American controversy.

But this narrative, with its strong focus on genetic engineering as a research tool, is strained. That the technique would move out of the research laboratory and into medicine, industry, agriculture and, most problematically, the military, was anticipated almost as soon as news of the first successful experiments began to circulate. As British molecular biologist Sydney Brenner wrote in 1974 to the British committee convened to examine genetic engineering, there were likely to be problems of controlling "institutions which can, and probably do, practise secrecy in their activities, such as defence research laboratories and, more importantly, the major drug companies".

This wider potential was marked by Stanford University's application for a patent on the early genetic-engineering techniques — an action that sent a shock wave through the NIH as the government sponsor of this work. "The patent" was widely perceived as a modestly seismic event, a nervous shift at the conjunction of the academic/not-for-profit and commercial tectonic plates sustaining the crust of the biomedical research enterprise," writes Fredrickson. From its inception, genetic engineering promised to yield socially disruptive technologies along with advances in scientific knowledge and socially useful products.

Fredrickson is aware of these larger dimensions of genetic engineering, yet they are never allowed to intrude on his narrative. During his tenure, he was in contact with

leaders of the pharmaceutical industry who made it clear that they would adhere to “the intent and spirit” of the NIH controls but not necessarily to their specific requirements — a position underscored by the flouting of the NIH controls by the rising star of the biotechnology field, Genentech, in 1978–79, when it ignored NIH procedures for approval of large-scale cultures of genetically modified organisms. That industry rebellion, underscoring the need for regulation of the new industry and posing a real dilemma for the NIH, which wanted at all costs to avoid legislation, remains invisible in this account.

Similarly, the prospect of military use of genetic engineering is mentioned but quickly dismissed on the grounds that the problem is addressed by the 1972 Biological Weapons Convention. That treaty has troubling loopholes, through which the genetic alteration of highly pathogenic agents of biological warfare has since slipped in the name of defence. These loopholes were not addressed when they were raised by members of Fredrickson's advisory committee in 1982; nor are these problematic developments confronted in this account.

Even viewing the genetic-engineering problem through the lens of research, Fredrickson skirts a question on which his narrative depends: did the risk-assessment experiments of the 1970s clear the field of significant hazard? An important answer is contained in a transcript of the first scientific meeting convened at the NIH in 1976 to discuss risk assessment. The transcript shows a key risk-assessment experiment being designed by the participants not to test a worst-case scenario, but rather to reassure the public that genetic engineering was harmless. As a participant noted, without any sign of moral indignation from his colleagues, “[This is] molecular politics, not molecular biology”. As an astute *Nature* reporter later observed of the treatment of the hazard question: “One must accentuate the positive. The evidence, however, does not seem substantial.”

Fredrickson acknowledges that the policies he pursued as NIH director, while yielding important scientific advances and medical benefits, also opened a Pandora's box of social and ethical problems. The new evidence on the recombinant DNA controversy promised by the release of his papers, and by the release of other sources as they become available, may reveal perspectives on the controversy more troubling than the soothing one presented in this memoir. ■

Susan Wright is a historian of science at Residential College, University of Michigan, East Quadrangle, Ann Arbor, Michigan 48109-1245, USA, and author of Molecular Politics: Developing American and British Regulatory Policy for Genetic Engineering, 1972–1982 (University of Chicago Press, 1994).

Mitigating mutations

The Cooperative Gene: How Mendel's Demon Explains the Evolution of Complex Beings/ Mendel's Demon: Gene Justice and the Complexity of Life

by Mark Ridley
Simon & Schuster/Weidenfeld: 2001/2000.
320 pp. \$26, £20

Andrew Berry

The Cooperative Gene is a masterpiece of scientific exposition. Mark Ridley has deftly packaged a hugely complex subject — the interplay of deleterious mutation with the evolution of genetic systems — into one long argument. He has distilled an area of biology that is rife with abstruse mathematics and the arcana of molecular genetics into a series of carefully explicated thought experiments and metaphors.

Despite the reassuring presence of a glossary, Ridley even manages to avoid jargon, and in principle the book is accessible to the general reader. However, it is not aimed at the Stephen Jay Gould-reading (or indeed Matt Ridley-reading) public. It lacks the anecdotal colour that is *de rigueur* these days if a book is to sail up the bestseller lists; there are neither pen portraits of crusty old professors peering down their microscopes nor strange tales of Y-linked ear hairiness in Indian men. Ridley's long argument is too dense and unrelenting to appeal to readers without a real interest in the subject; *The Cooperative Gene* is, in fact, a monograph masquerading as popular science.

Ridley's thesis is that the history of life has been shaped by evolution's attempts to overcome the effects of mutation. The occasional mutation is beneficial — indeed, they are the basis of adaptive evolution — but the vast bulk of all mutation is, without doubt, deleterious: “A random mistake in the DNA is about as likely to improve the creature it codes for as a random change in Hamlet is to improve the play.” Ridley's premise is that deleterious mutation must be kept within bounds if evolution is to occur: if the average rate of deleterious mutation exceeds one per genome per generation, then all the descendants of a parental generation with a mutation-free genome may be carrying deleterious mutations. Natural selection is then stymied because none of the competing genomes in the population is free from the taint of deleterious mutation and so it is not possible to select in favour of an intact, non-mutated genome.

Ridley sees a tension between the evolution of complexity — which necessarily involves encoding more information and therefore requires more genes — and the

depredations of mutation. Assuming a constant mutation rate per nucleotide, the large genome of a complex being is more likely to exceed the critical threshold of one deleterious mutation per genome per generation. Ridley concludes that: “Live complexity hits its ceiling when the DNA message is so long that a mistake happens every time it is copied. When life is near this upper limit, complexity cannot evolve upwards even if there is an ecological opportunity there.” The history of life has been punctuated by the evolution of “enabling mechanisms” that improve the fidelity with which hereditary information is replicated. The advent of DNA, which is less prone to degradation than RNA, as the vehicle for that information represents one such mechanism; the evolution of error-checking and DNA-repair systems is another; and sex yet another. With each innovation, the complexity ceiling rose.

Also militating against the evolution of complexity are genetic elements that subvert biological processes for their own ‘selfish’ ends, thus thwarting the cooperation between genes that is required for complexity. The simplest case of ‘mendelian law-breaking’ is meiotic drive, in which the probability of an allele being transmitted to the next generation is greater than the canonical mendelian expectation of 50%. For Ridley, the central player in the collision between the forces that promote complexity and those trying to tear it apart is meiosis. He emphasizes that the genetic randomization — through independent assortment and recombination — inherent in the mendelian process is critical in limiting the power and spread of genetic subversion.

Ridley's summary of the logic underlying this claim is masterful: “Mendelian justice is similar to the theory of human justice given by John Rawls in his book *A Theory of Justice*. Rawls argues that human beings will choose justice if they are operating behind a ‘veil of ignorance’. Imagine you are going to devise or apply laws, or allocate resources, among a number of people. Maybe you have to divide a pie among five people including yourself. If you know which piece is to be received by which individual you may allocate most or all of the pie to yourself; but if you do not (you are veiled in ignorance) you may divide the pie into equal pieces. The gene shuffling mechanisms we have been thinking about enforce Mendelian justice by drawing a veil of ignorance over the genes.”

Ridley refers to meiosis, the primary agent of mendelian justice, as “Mendel's Demon”, but the implicit parallel to science's other distinguished demon is more misleading than illuminating. In physicist James Maxwell's famous thought experiment, ‘Maxwell's Demon’ does the impossible: it oversees an imaginary process that breaks the laws of thermodynamics. In contrast, the process overseen by Mendel's is real: meiosis is no thought

experiment. In addition, whereas Maxwell's Demon acts as a non-randomizer, achieving its impossible goal by biased sorting of particles, Ridley's central message is that Mendel's acts as a randomizer (of genetic information). The two demons have nothing in common. Mendel would surely have had no truck with demons; Ridley should have followed his lead.

The Cooperative Gene is the latest product of what might be called the 'Oxford School' of evolutionary biology, whose current major protagonists include Richard Dawkins, Alan Grafen and David Haig. Starting from the inclusive-fitness arguments of W. D. Hamilton, the Oxford School takes a 'gene's-eye view' of evolutionary problems, and prefers the logic and elegant mathematics of game theory to the messy algebra of classical population genetics. But Ridley's book makes it clear that this approach is more than a mere alternative to traditional ones. Population genetics takes Mendel's laws as a given and applies them to populations, but Ridley is asking a more profound question: why and how did genetic systems, including meiosis, evolve? Thus population genetics essentially addresses the effects of the underlying causes in which Ridley is interested. *The Cooperative Gene* showcases a new way of thinking about evolutionary biology.

Ridley's logic is often beguiling, and he all too often commits the familiar Oxford School sin of being too easily satisfied by plausible speculation — if it looks kosher and sounds good, then it's good enough. Ironically, one of the few hypotheses championed by Ridley that is readily amenable to experimental testing — Alexey

Kondrashov's theory of the origin of sex — has recently indeed been tested, and found wanting. Kondrashov suggested that sex evolved to shuffle deleterious mutations, thereby facilitating the survival of individuals with relatively few mutations and the selective elimination of individuals with many. The theory requires that overall mutation rates in sexual species should therefore exceed the critical one-per-genome-per-generation threshold. However, a recent empirical study of mutation rates found that a substantial proportion of sexual species do not meet this criterion. Such troublesome facts notwithstanding, Ridley has written a marvellous book — one that brings the evolutionary analysis of genetic systems well and truly into the genome era. ■

Andrew Berry is at MCZ Laboratories, 26 Oxford Street, Cambridge, Massachusetts 02138, USA.

To begin at the beginning

Stem Cell Biology

edited by Daniel R. Marshak, Richard L. Gardner and David Gottlieb
Cold Spring Harbor Laboratory Press: 2001.
540 pp. \$115

Sean J. Morrison

Most tissues contain a small pool of undifferentiated cells called stem cells, which have the capacity to renew themselves by cell division, proliferate extensively and give rise

to many or all of the mature cell types found in the tissue. For example, embryonic stem cells of the inner cell mass of the early embryo give rise to all the cells of the body; haematopoietic stem cells in the bone marrow produce all blood cells; and neural stem cells produce different types of neurons and supporting glial cells. The current intense interest in stem cells lies in their potential to transform regenerative medicine, as well as the fact that recent advances have raised ethical issues related to the use of human fetal tissue as a source of stem cells.

The future of stem-cell biology will be to integrate what we know about stem cells in different tissues and at different times in development, so that we can both uncover common mechanisms of regulation and reveal the distinctions that guide stem cells to form different tissues. Yet most stem-cell biologists specialize in studying one type of stem cell from one tissue and have little basis on which to compare the properties of their cells with those from other systems.

Nonetheless, the similarities between stem cells in different systems have suffused the field with a belief that stem cells from different tissues will turn out to share many molecular properties. There is a widespread, but untested, expectation that different stem cells will use the same mechanisms to regulate common properties such as self-renewal and multipotency — the ability to form many different mature cell types. To test this expectation, we need to determine whether common genetic programmes are active in many different types of stem cells. Given the proclivity of stem-cell biologists to speculate unabashedly on the generality of their findings, some perspective is required to avoid myth-making.

Stem Cell Biology, edited by Daniel Marshak, Richard Gardner and David Gottlieb, will aid us in this task. This book considers the rich evolutionary and developmental complexity of stem cells in chapters ranging from yeasts to people, embryos to adults, and all three layers of the developing embryo — mesoderm, ectoderm and endoderm. The editors have done an excellent job of recruiting distinguished authors, including veteran leaders in the field such as Brigid Hogan and Austin Smith, and younger stars such as Markus Grompe. As a result, the book is strong on credibility and perspective. Each chapter is consistently written in a way that is accessible to the non-specialist. This is therefore a very good book both for newcomers to stem-cell biology and for experts looking for new ideas by broadening their perspective.

In addition to the requisite chapters on haematopoietic, neural, epidermal, embryonic and the connective-tissue-producing mesenchymal stem cells, the book includes a general consideration of progenitor-cell biology, including chapters on cell-type

Through the eye of the camera?

A photographic reconstruction (right) of Vermeer's *The Music Lesson* (below) reveals that a camera obscura may have been a part of the artist's armoury. *Vermeer's Camera: Uncovering the Truth Behind the Masterpieces* by Philip Steadman (Oxford University Press, £17.99, \$25) examines this controversial idea.



switching in yeast and cell-cycle control. Equivalence groups are also discussed; these are clusters of undifferentiated cells with equivalent developmental potentials from which individual cells are selected to adopt specific fates. This discussion of the fundamental mechanisms that control proliferation and differentiation in yeast, flies and worms was a very perceptive addition, for the mechanisms identified in these model systems resonate with the questions asked of mammalian stem cells in later chapters. For example, an early chapter discussing the general phenomenon of senescence provided an interesting perspective on later discussions of the ageing of liver and haematopoietic stem cells.

Two types of stem cell figure most prominently in the book: pluripotent stem cells — those able to give rise to all somatic cell types and germline cells — and haematopoietic stem cells. This is appropriate because, along with neural stem cells, they are the best characterized and most studied stem cells. Indeed, many of the hypotheses being tested in the context of liver and mesenchymal stem cells were conceived by analogy with haematopoietic stem cells.

Four chapters covering different aspects of the basic biology of haematopoietic stem cells give reasonably comprehensive coverage, although more systematic attention could have been paid to the different types of stem cells residing in different regions of the developing fetal haematopoietic system. These chapters provide an appreciation of the model on which much of stem-cell biology was conceptualized.

Similarly, an entire section of the book (six chapters) covers different aspects of stem cells in early development. They include a look at male and female germline stem cells in *Drosophila*, and use excellent illustrations to demystify the complex cell biology of each system. A particularly nice feature is Minx Fuller's side-by-side description of germline stem cells in flies and mammals. A chapter on trophoblast stem cells — cells that eventually form most of the placenta — and three chapters on different types of potentially pluripotent stem cells compose the rest of the section. Together, they illustrate the diversity of pluripotent cell types and the parallels between them. Given the current controversies over the derivation and use of human pluripotent cell lines, these chapters would be timely reading for non-specialists participating in the debate.

If the future of basic stem-cell biology depends on our ability to integrate our understanding of different stem-cell systems, and if the future of applied stem-cell biology depends on our ability to explain the field to non-specialists, then this book is a welcome step into the future. ■

Sean J. Morrison is in the Departments of Internal Medicine and Cell and Developmental Biology,

3215 CCGC, Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan 48109-0934, USA.

Reactions of a chemical nature

Nationalizing Science: Adolphe Wurtz and the Battle for French Chemistry

by Alan J. Rocke

MIT Press: 2001. 443 pp. \$42.95, £29.50

Georges Bram

A small street in Paris, and a street and a statue in Strasbourg, are now the only visible reminders of the distinguished nineteenth-century French chemist Charles-Adolphe Wurtz — the man who famously wrote that “chemistry is a French science”. Wurtz was a senator, a member of the French Academy of Sciences and a professor of organic chemistry at the Sorbonne. But he was not, as Alan Rocke explains, a ‘star’ of French science like Louis Pasteur, Marcellin Berthelot or Claude Bernard. Nevertheless, Wurtz’s death, on 12 May 1884, received extensive coverage in the newspapers, and his students and colleagues wrote obituaries and biographical memoirs soon after. Studies on various aspects of Wurtz’s work have been published more recently, but a comprehensive account was needed. This we now have from Alan Rocke in a book that will become a landmark in the study of the history of chemistry.

The author sets out his aim clearly — to provide a picture not only of the personal and intellectual life of Wurtz the man and scientist, but also of the social and political environment in which he lived, his daily working life and the evolution of scientific ideas.

Rocke gives us a good idea of the life and work of one of the major chemists of the nineteenth century. Many students passed through Wurtz’s laboratory. They came from France, particularly Alsace, Wurtz’s home region, and from other countries, and the lab was almost as important a school of chemistry as was the laboratory of Justus von Liebig in Giessen.

Wurtz was a tireless pioneer of the new theories that shaped organic chemistry. He promoted the ideas of Charles Gerhardt and Auguste Laurent who, in the first part of the nineteenth century, pioneered atomic theory in organic chemistry. And he later promoted the ideas of August Kekulé, who proposed the structure of the benzene ring. Completely convinced by atomic theory, Wurtz was opposed in this regard to many leading chemists such as Berthelot and Henri Sainte-Claire Deville, who belonged to the



French pioneer: Wurtz tirelessly promoted the new theories that shaped organic chemistry.

traditional school, which believed that the intimate structure of matter was not a legitimate question.

Rocke also describes chemistry at the time of Wurtz, and the chemists Wurtz knew. A chapter apiece is devoted to Liebig and Jean-Baptiste Dumas, Wurtz’s mentors. Another describes Berthelot, with whom Wurtz had a prolonged scientific fight over the question of atomic theory.

Rocke provides a valuable analysis of the decline of science in France in the second half of the nineteenth century — in particular, the decline of French chemistry in contrast to the rise of German chemistry. He discusses historians’ ideas on this question and proposes some explanations. These include the adverse effects of French university rules; the custom for academics to hold many positions at the same time, leading young scientists to miss out on job opportunities; the low salaries of young scientists; and the poor equipment in French laboratories. He gives figures that clearly show the French deficit in science students: in 1876, the total number of students studying in the faculties of science in French universities was 293; this was equal to the number of students working for a scientific PhD at Leipzig University alone. In spite of the high quality of many of its teachers, French chemistry was in decline.

One cause of the decline, as Rocke explains, was that, until the end of the nineteenth century, the atomic basis of organic chemistry was only taught in French science faculties if individual teachers decided to

teach it. Students who did not know their examiners' opinion on this matter would sometimes answer their examination papers using both 'languages' — that using atomic notation and that using the so-called equivalents in the formulae. Atomic theory was not introduced into French secondary schools until 1902 and was for a long time referred to as the 'atomic hypothesis'. This French reluctance to accept new theories was behind the late introduction of electronic theories in organic chemistry, which were not taught in French universities until the 1960s.

Until the late 1880s, France offered no higher technical education in chemistry. But in 1882, impressed by the quality of the engineers educated by the German Technische Hochschulen, one of Wurtz's students, Charles Lauth, founded the Ecole Municipale de Physique et de Chimie Industrielles in Paris, because he thought that France lacked managers for the chemical industry.

A final reason for the decline in French influence in chemistry was the poor patent policy — the product was protected, but not the process. The French chemical industry thus lagged well behind that in Germany, which had tight links with university research and drew strength from it.

In 1993, Rocke published *The Quiet Revolution: Hermann Kolbe and the Science of Organic Chemistry* (University of California Press), the German equivalent of this biography of Wurtz. Will he produce a useful equivalent for English science — perhaps *The Mauve Revolution: When Perkin made Chemistry Colourful*?

This review would not be complete without a comment on the (too?) famous quotation for which Wurtz is remembered. In 1868, he wrote: "La chimie est une science française. Elle fut constituée par Lavoisier d'immortelle mémoire" (Chemistry is a French science. It was founded by Lavoisier of immortal fame). Rocke points out, however, that these remarks, which frightened many scientists outside France, particularly those in Germany, were intended for a French audience. Wurtz wanted to stimulate the national spirit of French chemists and have them adopt the new structural theories of organic chemistry in a second revolution in chemistry (after that of Lavoisier) that he considered to be a consequence of the work of Laurent and Gerhardt.

I strongly recommend Rocke's book, not only to historians of science, but also to any scientist with an interest in the history of chemistry. I hope my chemist colleagues will not miss it and that the book will be widely bought by libraries in universities and research centres.

Georges Bram is in the Institut de Chimie Moléculaire d'Orsay and in the Groupe d'Histoire et de Diffusion des Sciences d'Orsay, Université de Paris-Sud (Orsay), Bâtiment 407, F-91400 Orsay, France.

Science in culture

Maintaining Masaccio

New data from the restored 'Trinity' in Florence.
Martin Kemp

Like cars and the human body, pictures need regular maintenance. This is particularly true of wall paintings, even when executed in the robust medium of fresco, in which the pigments chemically bond with an upper layer of moist plaster. Murals in publicly accessible spaces are particularly vulnerable to fluctuations in environmental conditions and accumulated dirt.

Masaccio's famous image of the *Trinity with the Virgin, St. John, Donors and a Skeleton* has enjoyed a particularly adventurous history. Painted in the mid-1420s in the church of Santa Maria Novella in Florence, it was the seminal demonstration of pictorial perspective in its earliest phase. But the fresco was totally covered up by an altarpiece 150 years later, when the interior of the church was completely remodelled. It was only rediscovered in the mid-nineteenth century, and its upper section was then transferred to the church's entrance wall.

It was finally returned to its original location and reunited with the battered and fragmentary lower portion containing the skeleton in 1950, a feat accomplished by Leonetto Tintori. Large areas of missing paint were restored; almost all the architectural structures in the lower portion were successively reconstructed, according to the best intuitions of how the ensemble originally worked.

Further work has recently been completed by Cristina Danti and her team from the Florentine conservation workshop Opificio delle Pietre Dure to ameliorate the inevitable changes and deterioration that have occurred over the half-century since Tintori's painstaking restoration. The Tintori and Danti campaigns, using traditional and technological means of examination, have together generated vast amounts of data about the highly technical optical construction of young Masaccio's illusionistic spaces (Masaccio died at the age of 27).

The key feature is what we now call the 'vanishing point', at which lines perpendicular to the plane of the picture appear to converge, and which Masaccio has placed at a reasonable height for an 'average' spectator. Even with the guidance of the converging parallels, the construction of the barrel vault is a far from trivial problem. Examination by Danti's team reveals that the left side of the curving vault is criss-crossed with a complex mesh of construction lines, some of which had been 'snapped' into the wet plaster with chords (which are pulled out and sharply released to leave an imprint), and others incised with a pointed instrument along a straight edge or with some form of compass. Having used the left side as his elaborate experimental field, Masaccio completed the right section with a greater economy of constructional effort.

A geometrical analysis of Masaccio's painted space — now confirmed by computer analysis —



reveals that he has intuitively adjusted its regular geometry to make it 'work' as a picture in terms of its actual site in the nave of the church. Among the many almost indiscernible subversions of canonical perspective are the circumferential extensions of the lowest row of coffers on left and right, presumably to make them appear to 'sit well' in relation to the arms of the cross. In addition to such empirical manipulations, the painter has insinuated some subtle asymmetries into the axial scheme. For example, God stands slightly off the central axis. We can also see that Christ's hands overlap the edges of his cross only on the left.

Although the prime, frontal viewing position was contrived to coincide with the width of the side aisle, the viewer entering from the usual entrance door in the fifteenth century, which was on the opposite side wall, would have seen Masaccio's illusion off-centre to his or her left. Masaccio has not exploited the full effect of parallax — which would have looked too extreme from other viewpoints — but he has subtly accommodated the asymmetrical line of approach.

The collective results arising from the old and new data indicate how the rules of construction and compositional intuition operate together in an experimental process that relies upon a constant interplay between geometry, judgement by eye, and supreme manual control.

Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

Celebrating science

Scientists feel passion, but they need to convey it in their writing, too.

John Carmody

In 1963, Sir Peter Medawar published a provocative essay, "Is the scientific paper a fraud?". With characteristic vigour, he asserted that the typical research paper misrepresents the reality of scientific work, in being written as if the venture unfolded in an unchecked and logical way. His argument has resonated with me for years, so I was delighted to discover that Robert Simmons feels the same way ("Sense and sensibility", *Nature* **411**, 243; 2001). In life, as in real laboratories and fieldwork, nothing proceeds with such smooth progress.

Perhaps more importantly, the "fraudulence" is not simply that the process is idealized; the passion is invariably drained too. This is the result of both the scientific culture and the reviewing process: important though both are, they etiolate the paper as an account of what the scientists have achieved. If (like art) science is concerned with understanding the world and our place in it, then it must be a passionate activity. Could anything be more important?

How much of that is apparent in scientific publications? If they have been stripped bare, do they falsify the nature of science? Contrast the cool brevity of Watson and Crick's 1953 account of the structure of DNA with the significance and impact of their paper about the molecule of the twentieth century.

The philosopher A. N. Whitehead thought that language is incomplete and fragmentary; is the problem of science that it must be made complete and concrete? In another essay, "Lucky Jim" (1968), Medawar asserted: "The history of science bores most scientists stiff." Possibly so, but the quiddity of what they do and how they account for that to themselves and their colleagues should be of deep concern. It is reassuring that such scientists as Lewis Thomas, Jared Diamond and Primo Levi (not to mention Medawar himself) have — when roaming beyond the fetters of the scientific paper — transcended that dichotomy.

Clinical scientists have a protean challenge. Their patients, rightly, want them to be highly competent but empathic communicators as well; their colleagues expect them to be simultaneously objective and involved. When they write, it is almost invariably as the dispassionate observer and detailed analyst, yet they can write with emotional honesty.

One, from an earlier era, was Herbert "Paddy" Moran. In his younger days he was a

great captain of the Australian Rugby XV; after postgraduate surgical studies, he worked with cancer patients, notably as a pioneer radiotherapist in Sydney. His professional publications are thoughtful but detached, and typical of the period. His three books of memoirs are quite different, especially the third, *In My Fashion* (1946), in which he described his reaction to the irony of his own diagnosis of melanoma and chronicled his decline towards death. It is intense writing.

"I observed that a pigmented mole which I had had upon the skin of my abdomen was breaking down ... I had no doubt, then, about the breaking down of this skin over a mole ... it is the fatal characteristic of this cancer that, almost at the moment when it appears locally to be breaking down, it has already scattered its evil seed at a distance." Soon after the pathologists confirmed the diagnosis he wrote: "There was nothing to do but to carry on" — he was a medical officer with the Australian Army in Britain — "The immediate effect on myself was a degree of shock from which I reached at once into bravado ... but very clearly I perceived that all my world had changed, whether I liked it or not ... I shall be often unhappy. I shall have pain, and pain, like many men, I never liked to bear ... I have my secret horror of being, before the end, weakened in body and in morale broken, that I shall whimper and complain."

In May 1945 he wrote: "The war ended in Europe yesterday. The crab was turning last night, dragging its tentacles. I awoke and was depressed. The nights are hardest to bear. I note in myself a tendency to slacken off all mental and physical activities. I must not so early throw in the sponge. I must summon up my will-power." A few days later: "The crab has been gnawing again. His nocturnal habits are disreputable ... The process is going to be more rapid than I had expected. Or is it that I am thinking too much on my little complaint? It is terribly hard not

Scientific literature should strive to achieve a sense that it is a chapter in an activity that has changed the world.



Herbert Moran's professional publications are detached, but his memoirs are quite different.

to think of it. Everything leads back to it."

Having decided to persist with the book, he reflected: "It does not matter if it never comes to birth in publication. It has saved me from a greater degree of unhappiness. After all, this death is a routine affair." By June it was: "My power of concentration is lessening. My interests are shrivelling. The will is like a tyre worn down to the tread. There is no excuse for this letting myself go." His sardonic humour did not desert him: speculating on the likely organ collapse, he wrote: "It is now ten to one on attrition through failure of the liver function. If only I had the energy to write a thriller: *Death Comes to the Evil Liver*."

Of course, scientific literature cannot be like Moran's eschatological prose. What it ought to strive to achieve — apart from the greatest lucidity — is a sense that it is a chapter in an activity that has changed the world and our insight into our relationship with it; a sense that the authors have been participants in an enormously important and successful human activity — one that they should be proud to celebrate. That would make the writing and the reading as thrilling as good science can really be. ■

John Carmody is at the School of Physiology and Pharmacology, University of New South Wales, New South Wales 2052, Australia.

Ultradivided matter

Pierre-Gilles de Gennes

Chemists in the nineteenth century recognized the existence of 'misbehaving' systems: hydrated alumina, starch, dextrin and gelatin, for example. Wolfgang Ostwald grouped them into a separate category, and coined the name 'colloids' (meaning 'glue-like'). Thomas Graham showed that the constituent objects had low mobility, and became convinced that they were aggregates of smaller molecules. They could be metal particles (such as in the gold colloids made by Faraday), oxides or organic systems.

This colloid concept was vague, at first even counterproductive, and delayed the birth of polymer science — for instance, rubber vulcanization was claimed not to be a chemical reaction (what we now call crosslinking) but rather a modification of an aggregated state. Ultimately, thanks to Hermann Staudinger, polymers were clearly identified and taken out of the colloid group. But colloids are clearly important: white paint, for instance, is a colloid based on titanium oxide and water. Most foods, cosmetics and inks are made up of colloidal structures.

As time went on, colloids became a well-defined family of objects — more properly defined as ultradivided matter. The family has many members: soap micelles, emulsions, suspensions of solids (with sizes from micrometres to nanometres), and newborn objects such as latex particles (which can be admirably tuned in composition and calibrated in size). A common feature is that the particles are floating in a solvent; if it is not too concentrated, a colloid can flow freely, which is convenient for industry.

Colloids differ fundamentally from solu-

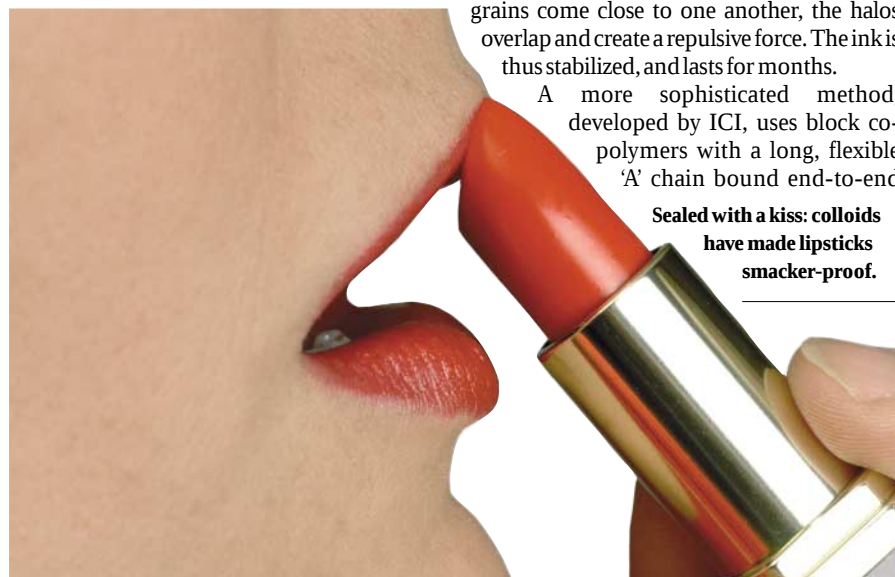
tions. In a solution, say of water and alcohol, all the relevant interactions (water–water, water–alcohol and alcohol–alcohol) are comparable to the thermal energy, kT . Thus, even if water 'prefers' water and so on, the tendency towards disorder — entropy — is dominant: the water–alcohol system does not segregate and, fortunately, remains a good solution. In contrast, ultradivided matter is unstable because of large interfacial energies, and the particle–particle interactions are stronger than kT . If the system is to remain fluid, the particles must not clump; the interactions have to be repulsive.

Neutral particles in a common solvent attract each other by Van der Waals forces, thus clumping together. How can this be avoided? Grains in water often carry an electric charge and repel each other by Coulomb forces. This is the stabilization process found by Faraday for gold colloids in pure water. The Dutch school of Verwey, Overbeck, Vrij and colleagues clarified this long ago. If salt is added, the Coulomb repulsions are screened out and the system clumps. Faraday detected this by a change of colour (from red to blue). Red corresponds to absorption of light by individual grains, blue to clumped systems.

It is often difficult to keep water salt-free on an industrial scale, so colloids must be protected by other means. The scribes of ancient Egypt needed ink, which was based on carbon black mixed with water. But this two-component mixture is unstable: the carbon grains attract each other by Van der Waals forces, clumping together and forming sediment within a few minutes. So the wise Egyptians added arabic gum (a polysaccharide obtained from the acacia tree). This is a long-chain polymer, which adsorbs onto the carbon grains, forming a 'halo' around each one. When two grains come close to one another, the halos overlap and create a repulsive force. The ink is thus stabilized, and lasts for months.

A more sophisticated method, developed by ICI, uses block copolymers with a long, flexible 'A' chain bound end-to-end

Sealed with a kiss: colloids have made lipsticks smacker-proof.



Colloids

These chemicals are important in many areas — most foods, paints, cosmetics and inks are made up of colloidal structures.

to another chain, 'B'. If the system is arranged so that A sticks to the grains, whereas B prefers the solvent, then a protecting halo of B is formed around each grain. This has been used in paints and is more efficient (but more expensive) than simple adsorption.

Two important industrial challenges have recently been solved by using suitable colloid systems. The first is to make water-based paints, thus avoiding toxic solvents, that give a painted water-resistant surface. One trick is to start from a standard suspension of negatively charged pigment (plus Na^+ ions to balance the charge) in which the grains are electrostatically stable. This is then deposited in conjunction with a solution of Ba^{2+} . The Ba^{2+} forms permanent bonds between two grains, and the deposited paint is thus water-stable. This role of multivalent ions is currently the subject of Byzantine discussions between theorists — but the technique works!

The second challenge is that lipsticks, which are based on pigments and waxes, must be soft when deposited but should not be transferable during a kiss. An efficient, non-transferring lipstick has recently been invented, incorporating a block co-polymer that is transformed by saliva into a robust gel.

There are many unanswered questions in the theory of colloids: for instance, some charged grains do not repel each other properly. This can be due to the presence of multivalent ions (as mentioned above), or to some local arrangement of the ions (at small inter-grain distances). This is important for biological molecules such as DNA and proteins. But colloid science is thriving, from cheap clays to sophisticated pharmaceutical systems, mainly because of the inventive minds of engineers, with the help of many available techniques (from expensive neutrons to cheap atomic force microscopes), and also with the modest assistance of theorists. ■

Pierre-Gilles de Gennes is at the Ecole Supérieure de Physique et de Chimies Industrielles, 10 Rue Vauquelin, 75231 Paris Cedex 05, France.

FURTHER READING

Morawetz, H. *Polymers: The Origin and Growth of a Science* (Wiley Interscience, New York, 1985).

Israelashvili, J. *Intermolecular and Surface Forces* (Academic, New York, 1985).

Cabane, B. & Henon, S. in *Collection de l'Ecole de Physique et Chimie* (Belin, Paris, 2001).

Proceedings of the 1st Lavoisier Seminar in *Collection de l'Ecole de Physique et Chimie* (Belin, Paris, 2001).

Caught in the crossflow

Elizabeth L. Brainerd

Suspension-feeding fishes filter water through complex structures in their throats. Food particles could clog the filters, but the fishes have a cunning system to prevent that happening.

Beverage manufacturers use a clever process known as crossflow filtration¹ to produce sparkling clear wines, beers and fruit juices. In this process, the fluid to be filtered flows parallel to the filter surface, rather than directly through it. By mechanisms that are not yet fully understood, any particles in the fluid tend to be transported away from the surface of the filter, minimizing clogging. Clear filtrate passes through the filter pores, and the particles are ejected from the system in a concentrated solution. On page 439 of this issue, Sanderson and colleagues² report that some fish species also use crossflow filtration — in this case to remove small food items, such as zooplankton and phytoplankton, from the water.

'Suspension-feeding' fishes have a filter-like set of structures called gill arches and gill rakers in the back of the throat (Fig. 1). Gill rakers have been assumed to function as a 'dead-end' filter, collecting small food items by sieving³ (Fig. 2a, particle 1, overleaf) or by hydrosol filtration, in which particles smaller than the size of the filter pores stick to the filter's elements^{4,5} (Fig. 2a, particle 2).

A potential problem with dead-end filtration, however, is that a lot of tiny food particles become trapped on the gill rakers. How can a fish remove the particles for swallowing without resuspending them and undoing all of the previous filtering? Some species solve this problem by using large amounts of mucus to trap particles on the gill rakers or on the roof of the mouth^{6,7}. Clumps of particle-laden mucus are then transported to the oesophagus. Other species, however, have little or no mucus on their oral surfaces⁸. For these species, as Sanderson *et al.*² show, crossflow filtration neatly avoids the problem because food particles remain in suspension rather than sticking to the gill rakers.

In a creative application of medical technology to zoological research, Sanderson *et al.* used a miniature fibre-optic endoscope to spy on the paths of food particles inside the mouths of three fish species: the gizzard shad, ngege tilapia and goldfish. They find that more than 95% of the particles travel towards the oesophagus without coming into contact with the gill rakers or any other oral surface. As water and suspended particles flow over the gill rakers, water exits between the gill rakers and food particles become more concentrated in the remaining water (Fig. 2b). A concentrated slurry of par-

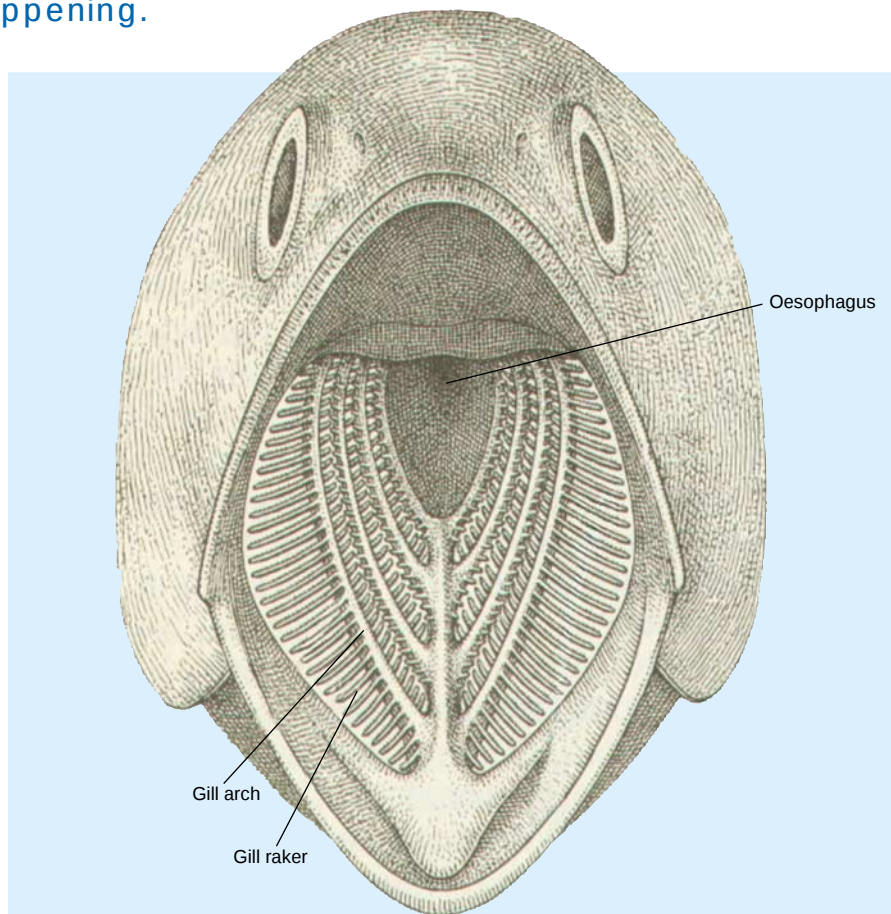


Figure 1 Suspension feeder. Gill arches and gill rakers form a filter-like surface at the back of the mouths of suspension-feeding fishes. Sanderson *et al.*² used a miniature endoscope to study the flow of water-suspended particles in three fish species: the gizzard shad, ngege tilapia and goldfish. The authors observe that more than 95% of food particles do not come into contact with the gill rakers but instead remain suspended as fluid flows parallel to the filter surface and water leaves between the gill rakers (crossflow filtration; Fig. 2b). Particles become more concentrated as they travel towards the oesophagus, and the fish swallows a concentrated slurry of food particles. (Figure reproduced from ref. 5.)

ticles accumulates in the back of the throat and is swallowed.

Neat though it is, this study leaves unanswered the question of precisely how crossflow filtration in fishes works. In fact, although crossflow filtration has been an important industrial process for decades, the underlying physical mechanisms here, too, are not yet fully understood. Why do particles tend to remain in the crossflow, rather than leaving with the filtrate or getting trapped in the filter pores?

During reverse osmosis — a type of crossflow filtration — simple diffusion seems to be sufficient to transport macromolecules and ions (less than 10^{-3} μm in

size) away from the filter surface¹. But fishes consume much larger particles (40–1,000 μm in size) and the mechanisms of particle transport for relatively large particles are more controversial. Using appropriate fluid-flow conditions, Sanderson *et al.*² tested one theoretical transport mechanism — radial inertial migration — and found that it is inadequate, by an order of magnitude, to explain particle transport in fishes that feed by crossflow filtration. Further theoretical developments in the field of filtration engineering, combined with more precise measurements of fluid flow inside the mouths of fishes, will be necessary to explain the physical mechanisms underlying cross-

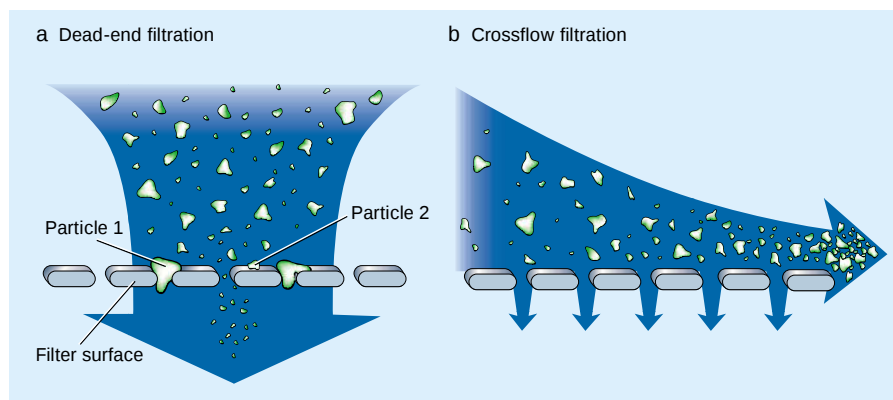


Figure 2 Dead-end and crossflow filtration. a, In dead-end filtration, fluid flow is perpendicular to the filter surface and the filter rapidly becomes clogged with particles. Particles may be retained by sieving when they are larger than the filter's pore size (particle 1), or by hydrosol filtration⁴ when they are smaller than the pore size (particle 2); in this case, the small particles stick to the elements of the filter. b, In crossflow filtration, fluid flows parallel to the filter surface and particles become more concentrated as filtrate leaves through the filter's pores.

flow filtration in suspension-feeding fishes.

Results from studies of suspension feeding have broad implications for ecological studies of freshwater and marine ecosystems⁹. Different suspension-feeding mechanisms select differently sized particles. Only particles larger than the filter pore are retained in a dead-end sieve (Fig. 2a, particle 1), whereas particles smaller than the pore size may be retained in dead-end hydrosol filtration (Fig. 2a, particle 2) or in crossflow filtration (Fig. 2b). The filtering mechanism determines which species and life stages of planktonic organisms are consumed, and this in turn has profound effects on the structure of aquatic populations and communities. The discovery by Sanderson and colleagues² of a fundamentally new mechanism of suspension feeding in fishes

may contribute to more realistic models of the dynamics of aquatic ecosystems. ■

Elizabeth L. Brainerd is in the Department of Biology and Program in Organismic and Evolutionary Biology, University of Massachusetts, Amherst, Massachusetts 01003, USA.
e-mail: brainerd@bio.umass.edu

1. Scott, K. & Hughes, R. (eds) *Industrial Membrane Separation Technology* (Blackie, London, 1996).
2. Sanderson, S. L., Cheer, A. Y., Goodrich, J. S., Graziano, J. D. & Callan, W. T. *Nature* **412**, 439–441 (2001).
3. Hoogenboezem, W., Lammens, E. H. R. R., MacGillavry, P. J. & Sibbing, F. A. *Can. J. Fish. Aquat. Sci.* **50**, 465–471 (1993).
4. Rubenstein, D. I. & Koehl, M. A. R. *Am. Nat.* **111**, 981–994 (1977).
5. Sanderson, S. L. & Wassersug, R. *Sci. Am.* **262**, 96–101 (1990).
6. Sanderson, S. L. *et al. J. Exp. Biol.* **199**, 1743–1756 (1996).
7. Sanderson, S. L., Cech, J. J. & Patterson, M. R. *Science* **251**, 1346–1348 (1991).
8. Goodrich, J. S., Sanderson, S. L., Batjakas, I. E. & Kaufman, L. S. *J. Fish. Biol.* **56**, 858–875 (2000).
9. Lazzaro, X. *Hydrobiology* **146**, 97–167 (1987).

Chemistry

How green was my ester

Giorgio Strukul

Hydrogen peroxide is an ideal oxidant. It cannot yet be used widely, because viable catalysts aren't available for many industrially important processes. But there are encouraging indications of progress.

Chemistry has turned green^{1,2}. The increased awareness of environmental problems has generated an overly simplistic division, however, especially in the media, between 'bad' chemistry — which first pollutes and then (sometimes) cleans up — and 'good', green chemistry. Chemists themselves are partly responsible for setting up this misleading contrast. But they are nonetheless among the leaders in trying to find less wasteful or damaging ways to handle the planet's resources. A lovely example appears on page 423 of this issue³, as described by Corma and colleagues.

Environmental protection and economic growth are not necessarily in antithesis, but improved chemical technologies can be needed to combine them. Industrial organic chemistry often involves multistep methods in which the yield of valuable end-product is largely outweighed by the amount of waste that has to be disposed of. The synthesis of complex molecules used as pharmaceutical or agrochemical products may involve up to 10–15 steps and can generate more than 100 kg of waste for each kilogram of end-product. In this respect, the use of catalysts that can simplify stepwise synthetic procedures,

or of cleaner reagents, is in principle more environmentally friendly.

The approach reported by Corma and colleagues³ fits the bill nicely. The authors have produced a new class of catalysts for the Baeyer–Villiger oxidation of ketones. This reaction, shown in Fig. 1a, was first described in 1899 by German chemists Adolf Baeyer and Victor Villiger⁴. It involves the transformation of a ketone into an ester (an organic molecule containing an extra oxygen atom with respect to the parent ketone), using an organic oxidant. Organic chemists have used the reaction for over a century. It is applied in the synthesis of a wide variety of chemicals, ranging from simple monomers used in the plastics industry to the more complex molecules that constitute pharmaceutical or agrochemical products⁵. However, as shown in Fig. 1a, the oxidant (generally an organic peroxy acid) produces waste — and often in larger amounts than the desired esters. So there has long been an interest in finding more environmentally friendly oxidants for this and similar reactions, hydrogen peroxide being the ideal choice.

Hydrogen peroxide, H₂O₂, is familiar because just about every household has a bottle of it, in dilute solution, for use as a disinfectant. Its main industrial applications are in the bleaching of paper, cellulose or textiles, or in making detergents; its applications in the manufacture of organic chemicals are minimal. The great advantage of hydrogen peroxide as a reagent is that it produces water as the waste product, and in small amounts only (Fig. 1b). The problem is that hydrogen peroxide does not react directly with substrates, but needs a catalyst to mediate the reaction. Over the years the search for such catalysts has progressed slowly^{6,7}, especially so in the case of the Baeyer–Villiger oxidation of ketones, for which the use of catalysts is relatively recent⁸. Unfortunately, the few catalysts capable of promoting hydrogen peroxide oxidation are almost exclusively metal complexes that are soluble in the reaction medium (homogeneous catalysts). These are not desirable for industrial purposes because they require several costly operations to separate the catalyst from the end-products.

Corma *et al.*³ open a new avenue for exploiting the Baeyer–Villiger process, based on insoluble (heterogeneous) catalysts that can be easily separated from the reaction medium by filtration. These catalysts are derivatives of a particular class of synthetic zeolite to which a metal — tin — has been added (Fig. 2). Zeolites are crystalline materials, many of which can be found in nature, that have a regular array of internal channels. They are composed of oxides of silicon and aluminium, and are similar to clays but with a different structure.

Incorporating tin atoms into the zeolite framework requires a special synthetic

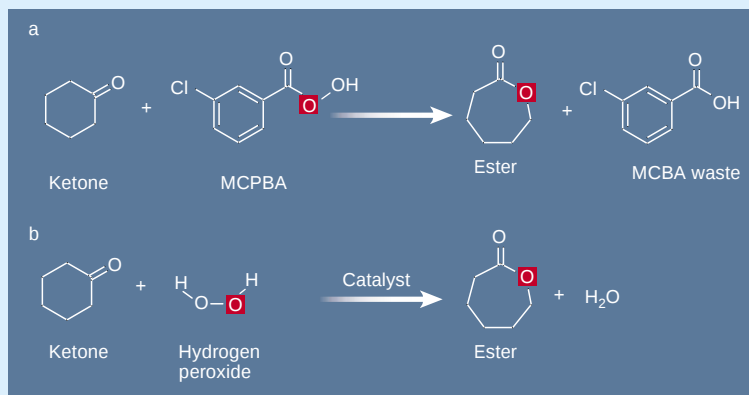


Figure 1 The Baeyer–Villiger oxidation of ketones. **a**, The conventional oxidation using a common organic oxidant (MCPBA, an organic peroxy acid). The reduced form of the oxidant (MCBA, an organic acid), is the waste product of this reaction. **b**, The same reaction, but using the zeolite catalyst of Corma *et al.*³ and hydrogen peroxide as the oxidant. Only a small amount of water is produced as waste.

procedure³. The tin centre is responsible for the activation of the ketone substrate and increases its reactivity in being oxidized by hydrogen peroxide. The new catalysts have exceptional selectivity. In the oxidation of complex ketones containing other, potentially oxidizable, functional groups, they oxidize only the ketone group — that is, they perform Baeyer–Villiger oxidation. The catalysts are insoluble in water and in all organic solvents, and the reaction occurs at the interface between the solid catalyst and the liquid

solution. At the end of the reaction, the catalyst can be removed and reused.

The catalysts devised by Corma *et al.* have been shown to work under laboratory conditions only, and moving from the production of just a few milligrams of ester to an industrial scale will not necessarily be straightforward. The catalysts will have to be synthesized in large quantities, their activity and lifetime will have to be improved, and their capacity to withstand industrial reaction conditions assessed. But they have

genuine potential to be of great benefit in ‘real world’ applications.

Zeolitic materials have already successfully catalysed other oxidation reactions involving hydrogen peroxide as the oxidant. For instance, the Italian company EniChem is using titanium silicalites⁹ to produce bulk chemicals such as propylene oxide, caprolactam and phenols. Titanium silicalites are a different class of zeolite, and are perhaps the most innovative oxidation catalysts to have emerged over the past 20 years. Given that background, one can be relatively optimistic about the prospects for the process developed by Corma *et al.* — it may prove to be a powerful tool in matching economic interests with sensible use of the environment. ■

Giorgio Strukul is in the Department of Chemistry, University of Venice, Dorsoduro 2137, 30123 Venezia, Italy.

e-mail: strukul@unive.it

1. Tundo, P. *et al.* *Pure Appl. Chem.* **72**, 1207–1228 (2000).
2. Anastas, P. T. & Warner, J. C. *Green Chemistry: Theory and Practice* (Oxford Univ. Press, 1998).
3. Corma, A., Nemeth, L. T., Renz, M. & Valencia, S. *Nature* **412**, 423–425 (2001).
4. Baeyer, A. & Villiger, V. *Ber. Dt. Chem. Ges.* **32**, 3625–3633 (1899).
5. Krow, G. C. *Org. React.* **43**, 251–798 (1993).
6. Strukul, G. (ed.) *Catalytic Oxidations with Hydrogen Peroxide as Oxidant* (Kluwer, Dordrecht, 1992).
7. Jones, C. W. *Applications of Hydrogen Peroxide and Derivatives* (R. Soc. Chem., Cambridge, 1999).
8. Strukul, G. *Angew. Chem. Int. Edn Engl.* **37**, 1198–1209 (1998).
9. Notari, B. *Adv. Catal.* **41**, 253–334 (1996).
10. http://www.iza-sc.ethz.ch/IZA-SC/Atlas/data/pictures/BEA_mod.html

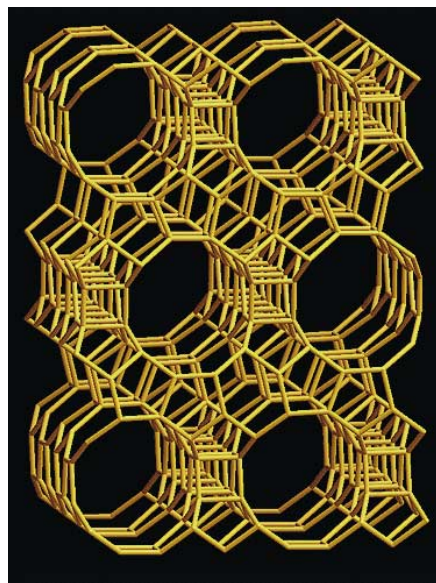


Figure 2 A β -type zeolite of the form used by Corma *et al.*³. This view of the silicon- and aluminium-oxide framework shows the array of internal channels in which catalysis takes place. In synthesizing the new catalysts, tin is substituted for some of the silicon or aluminium atoms facing the channel, and so is incorporated into the framework. Tin centres are responsible for the catalytic activity of these materials. (Reproduced from ref. 10.)

Neurobiology

Feeling bumps and holes

J. Randall Flanagan and Susan J. Lederman

We use our hands as well as our eyes to perceive physical objects. New insight into how our hands feel a surface may have implications for developing virtual-reality tools such as training devices for surgeons.

We can use our hands not only to manipulate the physical world, but also to perceive it. Using our hands to perceive the shape of an object often involves running the fingertips over the object's surface. During such ‘active touch’, we obtain both geometric and force cues about the object's shape: geometric cues are related to the path taken by the fingertips, and force cues to the contact forces exerted by the object on the fingertips. These cues are highly correlated, and it is difficult to determine the contribution of each to perception. On page 445 of this issue¹, however, Robles-De-La-Torre and Hayward describe an ingenious experiment in which they used a robotic device to uncouple force from geometric information. In a task involving the detection of bumps and holes, they show that force cues — not geometric cues —

determine perceived shape. The result has implications for virtual-reality applications, and may lead to new insights into how our perceptions of shape are built up from sensory signals obtained during active touch.

The use of one's hands to perceive the physical world is known as ‘haptic perception’. As a perceptual organ, the hand has several advantages over the eye: the hand can effectively ‘see’ around corners and can directly detect object properties such as hardness, temperature and weight. During active touch, the perceptual and motor functions of the hand are tightly linked, and people tailor their hand movements to the information they wish to extract². Whereas ‘local’ information about the object, at the fingertip's point of contact, can be extracted simply by touching the surface^{3,4}, more global features can be determined either by

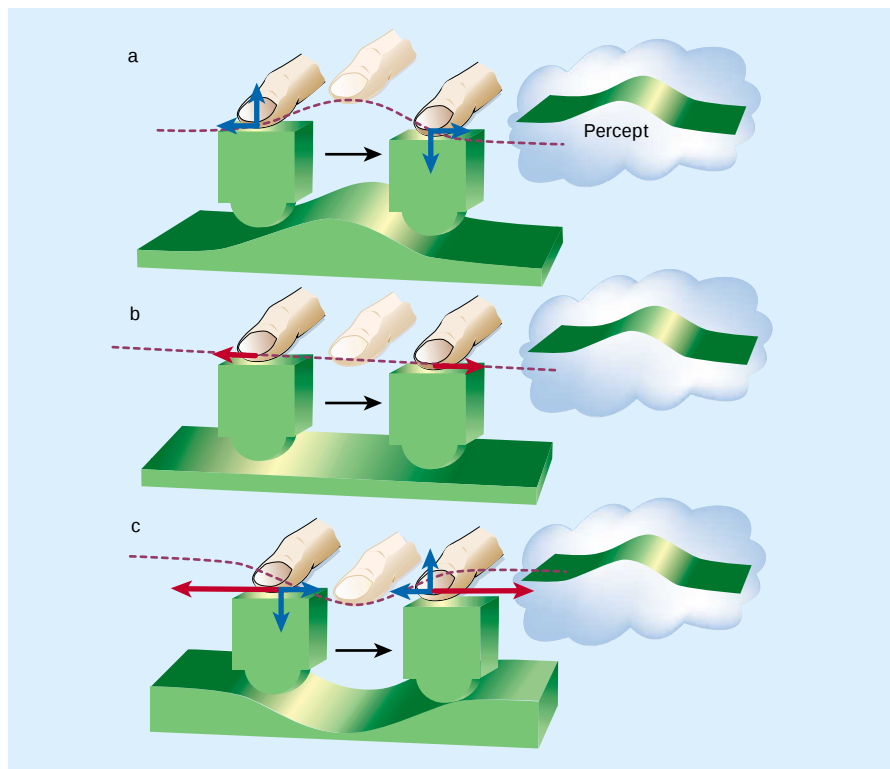


Figure 1 How force and geometric cues contribute to the perception of shape by touching. Robles-De-La-Torre and Hayward¹ asked people to use a fingertip to slide an object over a surface (which they could not see), and to indicate whether they perceived a bump or a hole. In all cases shown, subjects perceived a bump. **a**, The object traverses a real bump, which gives rise to the physical forces shown by blue arrows. Horizontal forces resist and then assist lateral motion as the object goes over the bump. Vertical forces cause the object and fingertip to rise and fall (dotted line; geometric cues). **b**, The object slides across a flat physical surface but horizontal virtual forces (red arrows) consistent with a physical bump are applied to the object through a robotic device. Although the fingertip does not move up and down, subjects perceive a bump. **c**, A virtual bump, twice the magnitude of that in **b**, is combined with a physical hole. The result is a stimulus that has the horizontal force properties of a bump but the vertical geometric properties of a hole. Although the fingertip falls and then rises with the object, subjects still perceive a bump.

enclosing the object in the hand or by moving the fingertips over the contours of the surface⁵ and integrating sensory inputs over time. Haptic information may also be obtained 'remotely' — using a tool — in a range of contexts, from neurosurgery to carpentry^{6,7}.

To study the role of force and geometric cues in haptic perception of shape, Robles-De-La-Torre and Hayward¹ asked people to slide an object, held beneath the index finger, across a horizontal surface and to indicate whether they perceived a bump or a hole. The surface could contain a real or a virtual bump or hole; the authors created the virtual bumps and holes by using a robotic device to manipulate the horizontal forces on the object and fingertip. When the object traverses a real bump, horizontal forces first resist and then assist the sideways motion of the object (Fig. 1a). At the same time, vertical forces drive the object and finger up and then down. The opposite occurs when a real hole is traversed. In these experiments, the virtual bumps and holes gave rise to the same horizontal forces as real bumps and holes, but without the concomitant vertical motion of object and fingertip (Fig. 1b). Subjects perceived the virtual features as bumps and holes even though the fingertip did not move vertically.

In a second experiment, the authors directly pitted force cues against geometric ones. When subjects experienced horizontal forces corresponding to a bump while tracing a real hole, they perceived a bump (Fig. 1c). Similarly, they perceived a hole when the horizontal forces corresponded to a hole, even when the finger moved up and down over a bump. The implication is that hori-

zontal forces provide the critical information needed to perceive shape by touch, and that these force cues dominate the geometric cues related to the path of the fingertip as it moves across a surface.

Practically, this work will impinge on the development of haptic interfaces used in virtual-reality applications. The results suggest that relatively inexpensive, planar 'force-feedback' devices could be used to create realistic perceptions of three-dimensional shape in a manner shown previously for virtual textures⁸. Ultimately, such devices may prove to be valuable in designing haptic communication systems such as virtual systems for surgical training⁹.

More broadly, the methodology has implications for fundamental neuroscience. Processing of sensory stimuli in primates, including humans, is extremely complex and involves numerous sensory-receptor systems and many different cortical and subcortical brain regions. The ability to uncouple force and geometric cues¹, together with functional magnetic resonance imaging of brain activity and neurophysiological analyses of neural processes, may help neuroscientists to unravel the neural circuits that underlie haptic shape perception.

There are, of course, some aspects of Robles-De-La-Torre and Hayward's results that are worth exploring further. Work on visually guided movement has shown that distinct neural pathways process visual information for action and for perception^{10,11}. Moreover, there have been several demonstrations that the motor system is not fooled by perceptual illusions^{12,13}. So one question is whether the perceptual effects

observed by Robles-de-la-Torre and Hayward translate into effects on action. For example, the shape of an object is one factor that affects how forces from the fingertips are controlled for stable grasping^{14,15}; it is not known whether virtual shape information obtained from force cues will also influence the control of grasping.

In addition, in the authors' experiments¹, the shape of a surface was sensed remotely, using a tool. We must be cautious in extrapolating their conclusions to tasks that involve direct touch, in which further sensory information is available. For example, during direct touch, the angle between the fingertip and the surface changes with the slope of the surface, and this may provide cues about shape¹⁶.

Nonetheless, the results reveal the importance of force cues in haptic perception of shapes. The next step will be to find out whether, and how, the contributions of force and geometric cues change with surface parameters such as the height and width of bumps and holes.

*J. Randall Flanagan and Susan J. Lederman are in the Department of Psychology, Queen's University, Kingston, Ontario K7L 3N6, Canada.
e-mail: flanagan@psyc.queensu.ca*

1. Robles-De-La-Torre, G. & Hayward, V. *Nature* **412**, 445–448 (2001).
2. Lederman, S. J. & Klatzky, R. L. *Cogn. Psychol.* **19**, 342–368 (1987).
3. Goodwin, A. W., John, K. T. & Marceglia, A. H. *Exp. Brain Res.* **86**, 663–672 (1991).
4. LaMotte, R. H. & Srinivasan, M. A. *J. Neurosci.* **7**, 1655–1671 (1987).
5. Louw, S., Kappers, A. M. L. & Koenderink, J. J. *Exp. Brain Res.* **132**, 369–374 (2000).
6. Brisen, A. J., Hsiao, S. S. & Johnson, K. O. *J. Neurophysiol.* **81**, 1548–1558 (1999).
7. LaMotte, R. H. *J. Neurophysiol.* **83**, 1777–1786 (2000).

8. Minsky, M. *Computation Haptics: The Sandpaper System for Synthesizing Texture for a Force Feedback Display*. Thesis, Mass. Inst. Technol. (1995).
9. Ottensmeyer, M. P., Ben-Ur, E. & Salisbury, J. K. *Stud. Health Technol. Inform.* **70**, 236–242 (2000).
10. Goodale, M. A. *et al. Nature* **349**, 154–156 (1991).
11. Milner, A. D. & Goodale, M. A. *The Visual Brain in Action* (Oxford Univ. Press, 1995).
12. Aglioti, S. *et al. Curr. Biol.* **5**, 679–685 (1995).
13. Flanagan, J. R. & Beltzner, M. A. *Nature Neurosci.* **3**, 737–741 (2000).
14. Wing, A. M. & Lederman, S. J. *J. Exp. Psychol. Hum. Percept. Perf.* **24**, 1571–1581 (1998).
15. Goodwin, A. W. *et al. J. Neurosci.* **18**, 10724–10734 (1998).
16. Jenmalm, P. & Johansson, R. S. *J. Neurosci.* **17**, 4486–4499 (1997).

Condensed-matter physics

In search of soft solutions

Douglas Durian and Haim Diamant

Physicists are turning their attention to delicate forms of matter, some of which appear mundane, but all of which are hard to understand. Fortunately, different materials share similar properties and problems.

Most forms of condensed matter are soft, but their physics is hard. Pick any apparently homogeneous material around you and wonder at the connection between its structure and behaviour: the magazine in your hands, the keyboard under your fingers, the flesh of your fingertips themselves, the contact lenses on your eyes, the cushion on your chair, the ink in your pen — the list is endless. These substances are composed of macromolecules and aggregates, with interactions that are too complex and too weak to form crystals spontaneously. Small external forces, slight perturbations in temperature, pressure or concentration, can all be enough to induce significant structural changes. Disorder, dynamics and deformation are the rule. Hence the phrase ‘soft condensed matter’.

The full diversity of this field was on display at the largest conference* ever hosted by the Center for Nonlinear Studies. Over 300 researchers gathered to discuss advances in understanding polymers, liquid crystals, biomolecules, surfactant solutions, membranes and vesicles, foams and emulsions, colloidal suspensions and granular materials. Each of these topics constitutes a rich field of research in its own right, but together they have evolved into an intellectually coherent branch of physics with a growing scope and promising future. Hans Frauenfelder, the director of the Center for Nonlinear Studies, underscored this in his welcome speech: “I believe soft matter, which goes all the way from glasses to living systems, will be the main area of work for years to come.” Indeed, at the conference it was clear that interest in glassy and biology-related systems is growing especially rapidly. More broadly, it emerged that many ideas cross over between topics, from the old to the new.

A well-established area of soft-matter research that continues to pose puzzles is the

self-assembly of surfactants in solution — an essential process in many natural phenomena, from the cleaning action of soap to the formation of biological cell membranes. Surfactants are molecules with hydrophilic head groups and hydrophobic tails. In aqueous solution, surfactants must solve a tricky problem: how to arrange themselves in such a way that their heads are exposed to water while their tails are shielded. Depending on various factors, possible structures include spherical micelles, flat lamellae and long, cylindrical ‘worm-like’ micelles.

There are long-standing questions about the flow properties of surfactant solutions.

When solutions of worm-like micelles undergo a non-uniform (shear) flow, their viscosity increases markedly, but no one knows why. The original idea, that the stressed micelles lengthen and entangle into a gel, can now be ruled out. One new proposal is that individual micelles can close up to form rings, and the size of these rings depends on the rate of shear. So at higher rates, rings are more likely to loop through one another, increasing the resistance to flow. Another idea is that short-range attractions cause long micelles to organize under stress into networks of bundles.

Whether or not bundling of worm-like micelles causes their increased viscosity, the hierarchical self-assembly of monomers into long filaments, bundles and higher levels of organization is ubiquitous in nature and is being widely studied. An important example is actin, whose self-assembled filaments serve as structural elements in muscular tissues and the cellular cytoskeleton. Depending on actin concentration and the presence of crosslinking proteins, the filaments form bundles or networks (see Fig. 1a), thereby strongly affecting the overall flow and elastic properties of the solution. Actin filaments can self-assemble even in a confined two-dimensional geometry when they associate with membranes. Another, less beneficial example is the self-assembly of misfolded proteins into filaments, ribbons and fibrils. The uncontrolled formation of these

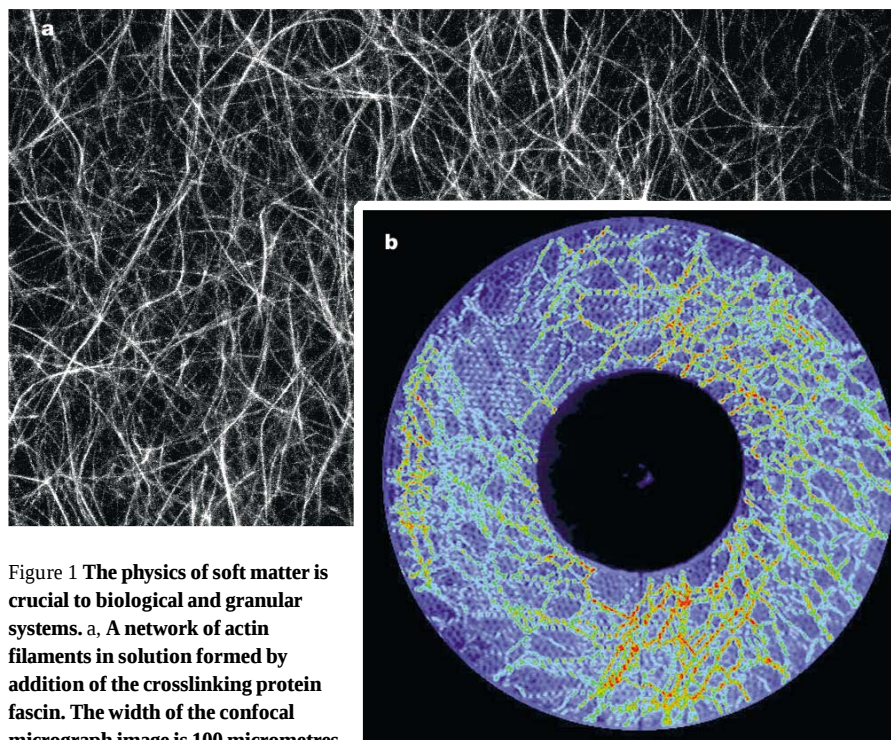


Figure 1 The physics of soft matter is crucial to biological and granular systems. a, A network of actin filaments in solution formed by addition of the crosslinking protein fascin. The width of the confocal micrograph image is 100 micrometres. (Picture courtesy of Yiider Tseng and Denis Wirtz.) b, The force pattern in a two-dimensional array of small plastic disks subjected to shear in a ring-shaped container. The stresses cause birefringence in the material, and in this photoelastic image red and blue regions correspond to strong and weak local forces, respectively. The outer diameter of the ring is 38.2 cm. (Picture courtesy of Robert P. Behringer.)

*Principles of Soft Matter, 21st Annual International Conference of the Center for Nonlinear Studies, Los Alamos National Laboratory, Santa Fe, New Mexico, USA, 21–25 May 2001.
<http://cnls.lanl.gov/Conferences/POSM>

structures is thought to be the cause of various diseases such as Alzheimer's.

The nonlinear interplay between structure and flow is important in many other forms of soft matter. For example, gas bubbles in foam or grains in a pile of sand remain completely at rest unless sufficiently large forces enable these packing units to unjam and flow around one another. This action is particularly chaotic for sand, with neighbouring grains colliding at a tremendous rate. It is crucial to study these collisions because they are inelastic and hence responsible for energy dissipation and the viscosity of the material. This is difficult because, by contrast with simple liquids, the collision rate is set by the act of flow itself, independent of the thermal energy of the system.

Not only is sand disordered and nonlinear, but the tools of statistical mechanics are seemingly useless in this context. So it is a relief that new experimental probes of granular fluctuations have become available, based on fast digital-video microscopy for surface grains and on intensity fluctuations of multiply scattered light for bulk grains. On the theoretical front, the idea of an 'effective temperature' based on the kinetic energy of colliding particles seems to offer a new and useful framework. It also suggests a deep analogy between granular materials and glassy liquids. As the actual or effective temperature is lowered, both types of system eventually jam — becoming stuck essentially forever in a restricted set of states. Once jammed, the mechanical properties depend on a network of force chains (Fig. 1b) that can be many grains long.

Jamming can also be induced as a function of density and external stress, with the results summarized by a 'jamming phase diagram', whose general features are common to all systems. Jamming also bears on the behaviour of biomolecular systems, such as the problems associated with protein folding. Here the packing units are the amino acids that make the protein sequence. To have biological function they should fold into a unique structure, yet proteins can jam in highly stable, wrongly folded configurations.

Soft condensed matter is arguably the most eclectic and interdisciplinary field of physical research. While clearly reflecting its diversity, the conference also vividly demonstrated a common research strategy, attempting to capture the essence of formidably complex systems using simple models and experiments. In this context, hierarchical structures in self-assembling systems, with biomolecules being a particularly rich example, and the unusual dynamics of glassy, jammed systems, were two of the most interesting themes to emerge.

Although there were no sessions on education, a frequent topic of informal discussion was the need to introduce physics students to soft matter earlier in their

careers. Participants bemoaned the lack of an all-encompassing textbook at the undergraduate level, and the want of a more appealing name than 'soft condensed matter'. Calling all would-be authors: now is the time.

Douglas Durian is in the Department of Physics and

Astronomy, University of California, Los Angeles, California 90095-154, USA.

e-mail: durian@physics.ucla.edu

Haim Diamant is at the James Franck Institute, The University of Chicago, Chicago, Illinois 60637, USA.

e-mail: diamant@control.uchicago.edu

Chemistry

Rings of destruction

Tomas Ganz

Many organisms use natural peptides to ward off microbes, but it's proved hard to make similar molecules for medical use. A ring-shaped peptide that might puncture microbial membranes could be the way forward.

It is known only too well that many infectious bacteria are becoming resistant to antibiotics¹. Multidrug-resistant strains of *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Enterococcus faecium*, among others, are now commonly encountered in hospitals and nursing homes. There is an increasing need for new drugs to fight these microorganisms. Yet, even though medicinal chemists have done a remarkable job of improving existing antibiotics, new antibiotic structures are all too rare. On page 452 of this issue², Fernandez-Lopez and colleagues report an exciting advance in this quest. They describe the design and initial testing of new antibiotics, based on a cyclic peptide structure, that selectively disrupt the cell membranes of target microorganisms.

Peptides, like the chemically similar proteins, consist of chains of amino acids, but unlike proteins they by definition have a relative molecular mass smaller than 10,000 (10K). Both animals and plants use natural antimicrobial peptides to defend themselves against bacteria and fungi^{3–5}. Nearly all such peptides are amphipathic and cationic; in other words they contain a hydrophobic face that interacts preferentially with lipids or other hydrophobic substances, and a positively charged hydrophilic face that interacts preferentially with water or with negatively charged surfaces. Driven by electrostatic attraction to the negatively charged cell membranes of microbial targets, these peptides enter the membranes, cluster, and form pores that eventually kill the microbes^{6–8}. The peptides spare membranes that contain mostly uncharged lipids, such as those of human and animal cells.

More than 200 natural antimicrobial peptides have been identified in plants, invertebrates and vertebrates, including humans. Many of these substances are active against bacteria and fungi that are resistant to existing conventional antibiotics. Why, then, have medically useful antimicrobial peptides, based on these natural molecules, been so slow to emerge?

One answer lies in the biological context in which these peptides function. Natural antimicrobial peptides are delivered to their targets with pinpoint accuracy by motile white blood cells that are either attracted to sites of infection or produced by epithelial cells (such as those that make up the surface of the skin or line the intestine) near the infection. A chemically synthesized version of the same peptide, if given by intravenous injection, would have to diffuse through

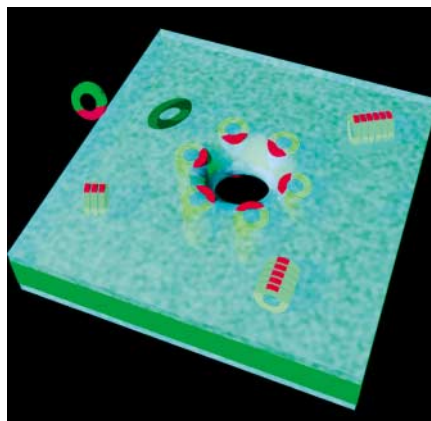


Figure 1 Fighting drug-resistant bacteria. As many bacteria are becoming resistant to existing antibiotics, there is a need to develop new types of antimicrobial compounds. Fernandez-Lopez *et al.*² have synthesized cyclic peptides that show a remarkable ability to puncture bacterial membranes and to kill bacteria *in vitro* and *in vivo*. The figure shows how these peptides might work^{6–8}. A peptide molecule is shown in the extracellular fluid above a section of a target membrane, with amphipathic cylinders in various stages of assembly around a completed pore. As more and more cylinders float in the membrane parallel to the surface, they introduce increasing strain into the membrane until the strain is partly relieved by a cluster of cylinders tipping over and forming a pore. Positively charged hydrophilic amino acids are in red, hydrophobic amino acids and membrane interiors in green, and negatively charged lipid head groups on membrane surfaces in cyan.

healthy tissues to reach the site of infection. Compared with conventional antibiotics, the relatively large — typically with a mass of more than 2K — and often highly positively charged antimicrobial peptides penetrate tissues slowly and are costly to produce. So initial applications of such peptides have been limited to topical preparations (creams, ointments and mouth rinses) that require small quantities of drug and less tissue penetration^{9,10}. Further improvements to these antibiotics are likely, but it is difficult to design them rationally because of the complexity of their interactions with membranes and with each other.

Clearly, it was desirable for researchers to extend their search beyond the common natural peptide structures. In living organisms, L-isoforn amino acids are the building blocks that make up peptides and proteins. The mirror-image D-amino acids are rare, especially in peptides of animal origin. Ghadiri *et al.*¹¹ previously synthesized rings consisting of six or eight alternating D- and L-amino acids; these rings included three consecutive hydrophilic amino acids and a stretch of alternating hydrophobic amino acids, L-tryptophan and D-leucine. In lipid membranes, the rings stacked into tubes that made the membranes permeable to molecules up to 10K in size.

The group's current study² was stimulated by the observation that one of these ring peptides showed antimicrobial activity towards several common bacteria. Starting with this peptide, the authors systematically replaced the three amino acids in the hydrophilic stretch with different amino acids. The substitutions altered the peptide's *in vitro* activity towards both *S. aureus* and another commonly encountered pathogenic bacterium, *Escherichia coli*, as well as its toxicity, as measured by its ability to damage red blood cells. Several of the peptides made bacterial membranes permeable and rapidly killed target bacteria.

On the basis of these *in vitro* results, Fernandez-Lopez *et al.*² chose three peptides for testing in mice that were infected with *S. aureus*. The peptides were very effective at clearing the infection, without substantial toxic side effects. Although the authors do not report detailed studies of how well these peptides penetrated the mouse tissues, it is encouraging that, in these preliminary experiments, the peptides were effective even when injected under the skin into tissue far from the site of infection.

The proposed mechanism of action for these peptides resembles that of magainins and other natural antimicrobial peptides that form structures known as α -helices in membranes⁷⁻⁹. These α -helical peptides fold up in the microbial membrane into their active shape, an amphipathic cylinder stabilized by hydrogen bonds. Fernandez-Lopez *et al.* suggest that their ring peptides stack to

form the same shape. In all cases, clusters of cylinders then form a pore through which small molecules can leak into and out of the microbe (Fig. 1).

One advantage of these new synthetic ring peptides is that they are much smaller than the α -helical ones, and this should increase their ability to penetrate tissues. But the presence of D-amino acids precludes the production of these peptides by available biosynthetic methods, and increases the cost of chemical synthesis — an important consideration given that these antibiotics might be required in amounts of the order of grams per dose. Non-peptide ring compounds with similar stacking and amphipathic properties might have some advantages over these peptides.

The development of a new antibiotic is a long and uncertain process. Many structural modifications have to be explored to increase activity, improve penetration and eliminate toxicity. Technological advances are

frequently necessary so that the candidate peptide can be produced at an acceptable cost. Extensive testing in animals and humans often brings up unforeseen problems. But, in the face of increasing microbial resistance to existing antibiotics, every new lead is welcome news. ■

Tomas Ganz is in the Department of Medicine, School of Medicine, University of California, Los Angeles, California 90095-1690, USA. e-mail: tganz@mednet.ucla.edu

1. Neu, H. C. *Science* 257, 1064–1073 (1992).
2. Fernandez-Lopez, S. *et al.* *Nature* 412, 452–455 (2001).
3. Lehrer, R. I., Ganz, T. & Selsted, M. E. *Cell* 64, 229–230 (1991).
4. Boman, H. G. *Cell* 65, 205–207 (1991).
5. Garcia-Olmedo, F., Molina, A., Alamillo, J. M. & Rodriguez-Palenzuela, P. *Biopolymers* 47, 479–491 (1998).
6. Huang, H. W. *Biochemistry* 39, 8347–8352 (2000).
7. Shai, Y. *Biochim. Biophys. Acta* 1462, 55–70 (1999).
8. Matsuzaki, K. *Biochim. Biophys. Acta* 1376, 391–400 (1998).
9. Ge, Y. *et al.* *Antimicrob. Agents Chemother.* 43, 782–788 (1999).
10. Mosca, D. A. *et al.* *Antimicrob. Agents Chemother.* 44, 1803–1808 (2000).
11. Ghadiri, M. R., Granja, J. R. & Buehler, L. K. *Nature* 369, 301–304 (1994).

Dyslexia

Talk of two theories

Franck Ramus

One possible reason why people with dyslexia have problems in learning to read is that some neuronal pathways involved in vision and hearing are damaged. That theory may need to be revised.

Regardless of how intelligent they are, people with developmental dyslexia have difficulty in learning to read, a characteristic first described more than a century ago¹. Dyslexia is now known to be a hereditary neurological disorder that affects a huge number of people — about 5% of the global population — but its underlying basis is still hotly debated. At two recent meetings*, however, the reasons for the disagreement became clearer.

The prevailing view of dyslexia involves the idea that learning an alphabetic writing system requires the brain to map letters to mental representations of the corresponding basic speech sounds (phonemes). The phonological-deficit hypothesis (Fig. 1a, overleaf) holds that people with dyslexia have specific problems in representing or recalling those sounds — hence the problems with mapping them onto letters. The theory is supported by observations that people with dyslexia have difficulty in retaining speech in short-term memory, and in consciously segmenting speech into phonemes (for example, deleting or substituting phonemes from words)².

But this view — that the cognitive basis of dyslexia is purely phonological — has been challenged by the discovery that people with this disorder also have an array of subtle sensory defects. For example, they fare less well than control subjects in several auditory tasks that require the perception of brief or rapid speech and non-speech sounds³. They also seem to have difficulties with several visual tasks, such as those that involve the perception of motion⁴. Moreover, there is evidence that the brains of some dyslexics have subtle neurological abnormalities in certain areas of the visual and auditory systems — the so-called magnocellular pathways⁵. Researchers who support the magnocellular theory (Fig. 1b) do not dispute the phonological-deficit hypothesis. Rather, they contend that phonological problems are caused by a basic deficiency in hearing sounds, and that a visual deficit might independently contribute to reading problems.

However, the magnocellular theory is itself now facing criticism. In several studies, auditory processing has not been found to be impaired^{6,7}. Moreover, although groups of people with dyslexia and control groups sometimes show significant differences in both visual tasks (J. Stein, J. Talcott, Univ. Oxford; K. Pammer, Univ. Newcastle) and auditory tasks (P. Tallal, Rutgers Univ.; C.

*The Fifth British Dyslexia Association International Conference, University of York, UK, 18–21 April 2001.
<http://www.bdainternationalconference.org/>
 Sensory Bases of Reading and Language Disorders, University of Essex, UK, 27–30 May 2001.
<http://www.essex.ac.uk/psychology/symposium2001/>



100 YEARS AGO

Only a few decades ago the real nature of tuberculosis was unknown to us: it was regarded as a consequence, as the expression, so to speak, of social misery, and, as this supposed cause could not be got rid of by simple means, people relied on the probable gradual improvement of social conditions, and did nothing. All this is altered now. We know that social misery does indeed go far to foster tuberculosis, but the real cause of the disease is a parasite — that is, a visible and palpable enemy, which we can pursue and annihilate... Such a conflict requires the cooperation of many, if possible of all, medical men, shoulder to shoulder with the State and the whole population; but now the moment when such cooperation is possible seems to have come... If we are continually guided in this enterprise by the spirit of genuine preventive medical science, if we utilise the experience gained in conflict with other pestilences, and aim, with clear recognition of the purpose and resolute avoidance of wrong roads, at striking the evil at its root, then the battle against tuberculosis, which has been so energetically begun, cannot fail to have a victorious issue.

Robert Koch

From *Nature* 25 July 1901.

50 YEARS AGO

The year 1950 was a good one for studying the movements of swifts (*Apus apus*), and during the year the movements of forty thousand birds were recorded: a report on the observations has been made by H. G. Hurrell and recently described in *British Birds* (44, No. 5; May 1951)... Swifts appear to penetrate the country from the south and work northwards. Arrivals over many years average a day or two earlier in the south than in the south-east. The east coast as a rule is reached rather late and usually in such small numbers that there is little to indicate any spring passage of swifts from the British Isles to other countries. It is difficult to say when the arrival period ends because movements which have a migratory appearance may take place at any time while swifts are in Britain. Large movements occur in June and early July. These are thought to be undertaken because unfavourable weather forces the swift to seek regions with more adequate food supplies; the food of the swift is adversely affected by the passage of a cyclone or depression.

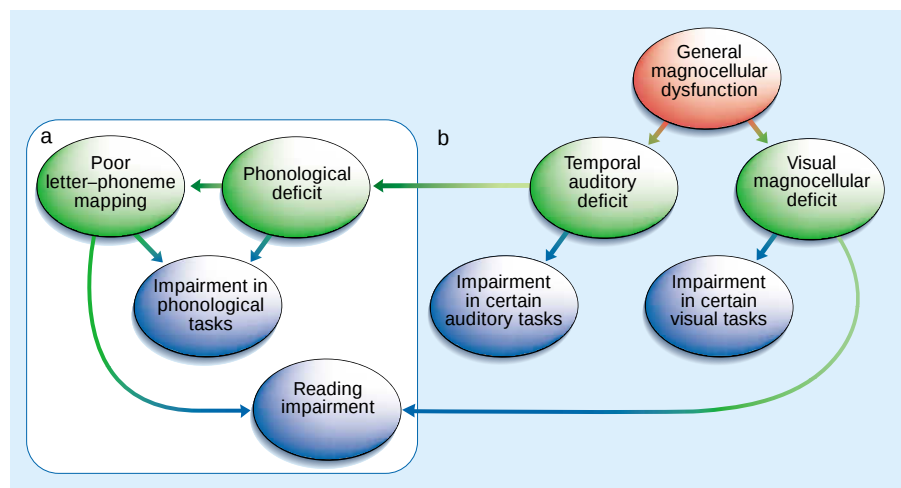
From *Nature* 28 July 1951.

Figure 1 The phonological and magnocellular explanations for dyslexia. Ovals represent impairments at the neurological (red), cognitive (green) or behavioural (blue) levels; arrows represent causal connections. **a**, The phonological theory holds that the core cognitive deficit lies in the ability to represent or recall speech sounds (phonological representations). This results in defects in mentally mapping letters to phonemes, and leads to reading impairments and problems with phonological tasks. The neurological basis for the phonological deficit, however, is not yet known. **b**, The magnocellular theory is based on the division of the visual system into two neuronal pathways: the magnocellular and parvocellular pathways. This theory holds that the magnocellular system is abnormal in people with dyslexia, causing difficulties in some aspects of visual perception and in binocular control that may cause a reading impairment. In addition, similar impairments in the auditory system are suggested to cause a deficit in processing the rapid temporal properties of sounds, leading to the phonological deficit.

Witton, Univ. Oxford), it is becoming apparent that these group effects result from only a minority (typically one-third) of the dyslexic participants. The remaining two-thirds fare normally (S. Rosen, Univ. College London). Other teams have failed to find significant differences between dyslexic and control groups, but usually discover a few people in the dyslexic group who do have sensory defects (Y. Griffiths, M. Snowling, Univ. York; S. Heath, Univ. Western Australia; M. Van Ingelghem, KU Leuven; S. Amitay, Hebrew Univ. Jerusalem). So, although most studies have shown that some dyslexic people have sensory deficits, the prevalence and significance of such deficits remain uncertain.

One consideration is the tasks used to investigate these defects. Given that attention and general cognitive abilities are involved in sensory tasks, it is not surprising that differences in auditory and visual processing can sometimes be explained by differences in non-verbal intelligence (M. Ahissar, K. Banai, Hebrew Univ. Jerusalem). Likewise, verbal skills, such as verbal short-term memory, may influence performance. A typical task requires a person to judge the order of events: two, three or four stimuli are presented one after the other, and the subject must recall the order of the stimuli, or say which one was different from the others.

Clearly, this task requires storing the stimuli in short-term memory, which might involve a verbal strategy; subjects might mentally rehearse the words 'high, low' so

as to remember that they heard a high tone followed by a low tone. Yet many people with dyslexia also have problems with verbal short-term memory — a characteristic of the phonological deficit. Indeed, when people are explicitly instructed to use a verbal strategy in this task, control subjects fare better than before, whereas dyslexics perform more poorly (C. Marshall, Univ. York). Similarly, dyslexics have difficulty in discriminating between some visual stimuli when presented sequentially, but not when presented simultaneously (G. Ben Yehuda, Hebrew Univ. Jerusalem). Perhaps, then, the presumed sensory deficits actually reflect strategic differences.

Even when sensory deficits are found, they do not always appear in the ways predicted by the magnocellular theory. For instance, the magnocellular visual pathway is most sensitive to stimuli of low spatial frequencies, presented under low lighting and at low contrast. So one would predict visual defects in dyslexics to be most apparent under these conditions (for example, when having to detect gratings of thick light-grey and dark-grey stripes appearing on a screen in the dark). Yet people with dyslexia typically fare worse than control subjects at all spatial frequencies (that is, whether stripes are thin or thick)⁸ (G. Ben Yehuda), as well as at tasks in which frequencies are not entirely controlled and contrast and lighting are high (M. Bradshaw, Univ. Surrey; M. Van Ingelghem). Similarly, the theory predicts that the auditory impairment will be limited

to rapid or brief sounds, but dyslexics sometimes fare normally when sound frequencies change rapidly^{9,10}, and badly when they change slowly¹¹. At best, the implication is that the magnocellular theory needs some revision.

More worrying is that there is little evidence for a link between these simple sensory deficits and phonology and reading. Indeed, no magnocellular deficit has been found in people who have a reading impairment as a result of eye strain or visual distortion (A. Wilkins, Univ. Essex). Similarly, performance in auditory tasks does not seem to predict performance in speech processing (C. Watson, Indiana Univ.; C. Marshall). Moreover, a study of twins showed that phonological impairments are highly heritable, whereas auditory deficits are more likely to be due to environmental factors¹². Of course, as the auditory system is a major source of input to the phonological system, any auditory defects are likely to affect phonological processing (M. Merzenich, Univ. California, San Francisco). But the question here is whether the auditory deficit suffered by some dyslexics is of the type and magnitude to cause the kind of phonological deficit that in turn causes a reading impairment.

Could there be a hidden factor that might explain why the incidence of sensory deficits varies so much between studies? One factor might be the occurrence of other developmental disorders in the dyslexic population — many dyslexics also have specific-language impairment or attention-deficit disorder. There is a growing suspicion that sensory deficits might be found only in these people⁶ (G. McArthur, Univ. Oxford; my own data). The implication is that groups of people with 'pure' dyslexia might be most relevant to understanding the causes of the reading impairment. Finally, the phonological-deficit theory also requires further investigation. Indeed, the greatest challenge is still to try to understand the precise nature of the phonological deficit and its biological cause.

Franck Ramus is at the Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK.

e-mail: f.ramus@ucl.ac.uk

1. Pringle-Morgan, W. *Br. Med. J.* **2**, 1378 (1896).
2. Snowling, M. J. *Dyslexia* (Blackwell, Oxford, 2000).
3. Tallal, P., Miller, S. & Fitch, R. H. *Ann. NY Acad. Sci.* **682**, 27–47 (1993).
4. Stein, J. & Walsh, V. *Trends Neurosci.* **20**, 147–152 (1997).
5. Galaburda, A. & Livingstone, M. *Ann. NY Acad. Sci.* **682**, 70–82 (1993).
6. Heath, S. M., Hogben, J. H. & Clark, C. D. *J. Child Psychol. Psychiatry* **40**, 637–647 (1999).
7. Hill, N. I., Bailey, P. J., Griffiths, Y. M. & Snowling, M. J. *J. Acoust. Soc. Am.* **106**, L53–L58 (1999).
8. Skottun, B. C. *Vision Res.* **40**, 111–127 (2000).
9. Mody, M., Studdert-Kennedy, M. & Brady, S. J. *Exp. Child Psychol.* **64**, 199–231 (1997).
10. Rosen, S. & Manganari, E. *J. Speech Lang. Hear. Res.* (in the press).
11. Witton, C. *et al. Curr. Biol.* **8**, 791–797 (1998).
12. Bishop, D. V. *et al. J. Speech Lang. Hear. Res.* **42**, 155–168 (1999).

Planetary science

Uncool Callisto

Kristin A. Bennett

Models of heat convection suggested that any liquid on Callisto, one of Jupiter's moons, must be frozen. But those models did not take into account the different properties a surface layer of ice might have.

Jupiter's four largest moons, Io, Europa, Ganymede and Callisto, are known as the galilean satellites. These are dark and mostly dynamic worlds, and their present state and evolution can usefully be compared with Earth. For various reasons, Callisto has been seen as the least interesting of the four. But this may be no ugly duckling among the moons. As Ruiz points out on page 409 of this issue¹, if the different properties of ice under various conditions are taken into account, there may be a liquid ocean 150 km beneath Callisto's icy surface.

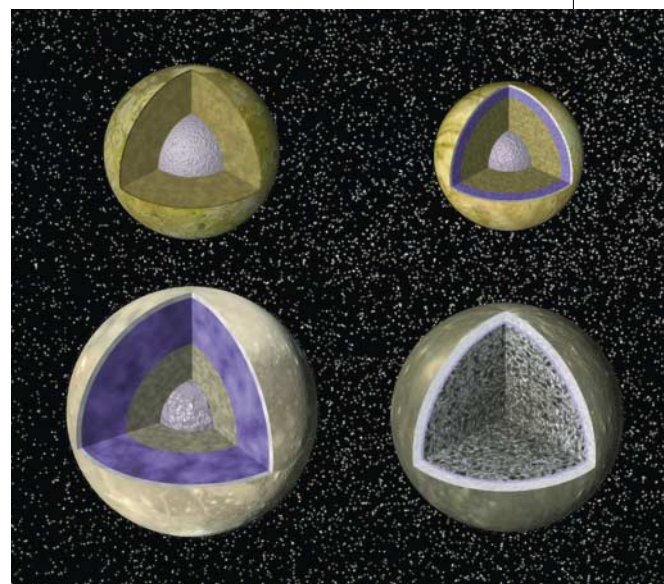
The galilean satellites are vast: Callisto, for instance, is the size of Mercury. And like planets they are thought to consist mainly of ice and rock, more or less differentiated into inner cores, convecting mantles and outer rigid crusts (Fig. 1), and are active to varying degrees. Of the four, Callisto has always been the odd one out. It seems to be a barely differentiated solid ball of ice and rock, covered by a rigid rim of ice. Around 80% of its surface is cratered and pitted, from which it is inferred that, unlike its companions, Callisto has been volcanically and tectonically dead almost since its origin². Frankly, from a geologist's point of view, Callisto was boring. But Ruiz shows that an internal water-ice ocean ebbing and flowing under its surface is perfectly plausible.

Ruiz's theme, as stated in the title of his paper, is the "stability against freezing of an internal liquid-water ocean in Callisto". Magnetometer data³ from the Galileo space-

craft show that Callisto has a variable magnetic field, which is most easily explained by a dynamo effect stemming from electronic conductivity of an electrolyte — water with a high salt content. But all other evidence suggested that any water would be frozen, because the heat to keep it warm would have long since been lost through convection. Ruiz's rigorous geophysical model, based on the properties of different forms of ice and water, shows that there can indeed be an ocean in Callisto's depths, even without a high salt content in the water to lower the freezing point. He calculates that a 'stagnant' outer ice shell on the moon would remain resistant to internal heat loss. The implication is that Callisto may have retained an ancient liquid ocean, now some 20 km thick.

To keep an ancient liquid-water ocean from freezing inside a body the size of Callisto, there must be both a heat source and a mechanism for impeding heat flow. Callisto is too far from the Sun to gain any solar heating, and there is no 'tidal heating' from the energy generated in a surface ocean by its orbit around Jupiter. So all heat is assumed to be generated by the decay of radioactive elements. One way to maintain a liquid ocean is by an antifreeze substance that reduces the freezing point of water ice. Another way would be if the outer ice layers are more rigid and resistant to heat flow out through them than has been assumed. Heat can be transferred through conduction or convection, and there have been numerous attempts with

Figure 1 The galilean satellites. Clockwise from lower left: Ganymede, Io, Europa and Callisto. The surfaces are mosaics of images obtained in 1979 by the Voyager mission, and the interior characteristics are inferred from gravity- and magnetic-field measurements by the Galileo spacecraft. The satellites are shown according to their relative sizes (Callisto's radius is 2,403 km). All except Callisto have metallic cores of iron and nickel. Image taken from ref. 8.



NASA/JPL

thermal models to account for the data on the differentiated moons obtained from the Galileo spacecraft^{4,5}. Later models have been more sophisticated⁶, but have not considered the physical properties of the material being modelled.

Ruiz does so. He takes account of the fact that ice is highly anisotropic, and that its structure changes dramatically under extreme temperatures, pressures and stress conditions. There are over 12 known different crystallographic structures of ice, as well as two amorphous (non-crystal) types and of course the liquid form, and each is known to exist in the Solar System. Scientists have determined the viscosity (response to stress) of ice at temperatures, pressures and stresses relevant to the icy moons⁷, and have shown that its viscosity varies drastically with pressure and temperature. But all previous models of ice on Callisto have taken the viscosity of ice to be that of water. Ruiz not only invokes the experimentally established viscosity of water ice found on the surface of Callisto, but also considers differing microstructural conditions, such as ice-grain mobility and sliding during convection.

Water seems simple, yet it is one of the most complicated molecules in the Solar System. By including the structural response of water and ice in models of the thermal

evolution of a planetary body, Ruiz has made a leap in planetary modelling. His analysis brings into question our understanding of ice and water in the outer Solar System, and — just for starters — will force a re-evaluation of the thermal and structural models of the largest 14 or so moons of Jupiter.

If Ruiz is right, and there is a liquid-water ocean flowing and driving currents deep within Callisto, then we might be closer to solving the puzzle of why the sister moons Ganymede and Callisto are so similar in size and location, yet so opposite in nature. And of course, where there is any suggestion of water elsewhere in the Solar System other than Earth, there are suggestions of the possible existence of life. ■

Kristin A. Bennett is at Los Alamos National Laboratory, Los Alamos Neutron Science Center, MS H805, Los Alamos, New Mexico 87545, USA. e-mail: bennett@lanl.gov

1. Ruiz, J. *Nature* **412**, 409–411 (2001).
2. Cassen, P. M., Peale, S. J. & Reynolds, R. T. *Satellites of Jupiter* (Univ. Arizona Press, Tucson, 1982).
3. Zimmer, C., Khurana, K. K. & Kivelson, M. G. *Icarus* **147**, 329–347 (2000).
4. Lewis, J. S. *Science* **280**, 1573–1576 (1998).
5. Consolmagno, G. J. & Lewis, J. S. *Icarus* **34**, 280–293 (1978).
6. Solomatov, V. S. *Phys. Fluids* **7**, 266–274 (1995).
7. Durham, W. B., Kirby, S. H. & Stern, L. A. *J. Geophys. Res.* **102**, 16293–16302 (1997).
8. <http://photojournal.jpl.nasa.gov/cgi-bin/PIA01082>

Population ecology

Birds in a buffer state

There is a happy history of long-term data collected by volunteer ornithologists being used to tackle ecological questions. The latest example is described elsewhere in this issue (*Nature* **412**, 436–438; 2001) by Jennifer Gill and colleagues. Their subject is the black-tailed godwit, *Limosa limosa* (pictured), a large migratory wader.

Godwits that breed in western Europe, including Britain, mostly winter in Africa. But birds from an Icelandic population winter around the British coast, returning to Iceland each spring to breed. For reasons that are unclear, this Icelandic population has increased steadily for several decades, resulting in more winter migrants to Britain.

Combining census data and new surveys, Gill *et al.* find that British estuaries vary in quality as godwit habitats. Bird numbers at the best estuaries



have remained relatively stable, and the overall population increase has been buffered by a disproportionate increase in birds using lower-quality sites. This cannot be explained by improved habitat quality, as prey availability does not seem to have changed. Gill *et al.* find that the food-intake rate of godwits at poorer sites is lower than on the traditionally favoured estuaries.

Consequently, these birds have higher winter mortality; and those that do survive reach the

Icelandic breeding grounds later, perhaps missing out on the best breeding territories.

Lower probability of survival or reproductive success could be a mechanism for population regulation. When the total population density is low, high-quality sites will be favoured. But as their numbers increase, some birds must use lower-quality habitats where they fare less well. Such density-dependent regulation could eventually check the current population increase. **Rory Howlett**

Daedalus

Deft definitions

All language depends on agreed definitions of its words. For public objects and actions this is easy; we can point to what a word refers to. But words for private sensations cause endless confusion. Defining them needs a subtler approach.

Daedalus has been inspired by a recent study in which volunteers claiming to be 'in love' were placed in a magnetic-resonance brain scanner, and shown photographs of their alleged beloved. Four specific brain regions immediately gained activity, and one lost it. Photographs of mere friends gave a different pattern.

The word 'love' must be responsible for more confusion and misery than any other in the language. How are we supposed to learn the difference between desiring, fancying, being infatuated with, falling in love with, being in love with, or loving, somebody? Do these states exhaust the emotional range? Are they the same for men and women? And what are the logical connections between them? (Daedalus reckons, for example, that falling in love leads to the state of infatuation, and that you can get to be in love with somebody without having fallen in love with him or her.) Romantic fiction merely confuses the issue. It generates innumerable emotional agonies in those who take it seriously.

So DREADCO's psychologists are now launching a programme of emotional exploration. Volunteers from starry-eyed teenagers to ruthless Casanovas to long-married pensioners are being studied by all brain-scan technologies. At first the programme will be merely empirical. It will seek to correlate the brain maps of volunteers in various emotional states with what they claim to be feeling. In due course the team will discover the most consistent use of the existing words. New words, or new definitions, may have to be invented to fit the observed facts. The baffling private world of love will slowly gain an objective public map.

The programme will then switch into its educational phase. People confused by their emotional state will take a DREADCO brain scan. They will learn the right expression for their state of mind, and be taught how to recognize other states. The DREADCO standard vocabulary and definitions will spread through society, ending the anguished doubts and misunderstandings propagated by the current versions. The question 'Do you love me?' will at last have a true, demonstrable answer. **David Jones**

Chemistry beyond the molecule

Gautam R. Desiraju

Supramolecular chemistry has grown in importance because it goes beyond the molecule — the focus of classical chemistry. It also offers a fresh interface with biological and materials science.

For a long time chemists tried to understand nature at a level that was purely molecular — they considered only structures and functions involving strong covalent bonds. But some of the most important biological phenomena do not involve making and breaking covalent bonds — the linkages that connect atoms to form molecules. Instead, biological structures are usually made from loose aggregates that are held together by weak, non-covalent interactions. Because of their dynamic nature, these interactions are responsible for most of the processes occurring in living systems. Chemists have been slow to recognize the enormous variety — in terms of structure, properties and functions — offered by this more relaxed approach to making chemical compounds.

The slow shift towards this new approach began in 1894, when Emil Fischer proposed that an enzyme interacts with its substrate as a key does with its lock¹. This elegant mechanism contains the two main tenets of what would become a new subject, supramolecular chemistry^{2,3}. These two principles are molecular recognition and supramolecular function.

Molecular recognition is implicit in the lock-and-key model — provided both the geometry and the non-covalent interactions are compatible between the interacting partners, you get recognition. Such highly specific interactions also lead to useful supramolecular functions. For example, it is important that an enzyme works only on the appropri-

ate substrate. A key without its own lock or a lock without its own key is quite useless.

The initial motivation behind supramolecular chemistry was to design chemical systems that mimic biological processes. The rise of the supramolecular approach was aided by observations of stable compounds that did not involve covalent bonds. Early examples of these 'addition products' include donor-acceptor complexes and clathrate compounds (Fig. 1). Some donor-acceptor complexes do not involve normal covalent bonding. Instead, they are held together by one molecule donating electrons, or perhaps sharing a hydrogen atom, with another.

A classic example of a donor-acceptor complex is formed by silver ions (Ag^+) and ethene ($\text{CH}_2=\text{CH}_2$), in which the ethene donates some electrons from its double bond to Ag^+ (Fig. 1a). The interaction is not so strong that it leads to a covalent bond, but it is strong enough to form a stable complex.

Back in 1948, H. M. Powell⁴ described a series of what he called clathrates — derived from the Latin *clathratus*, meaning 'enclosed by the bars of a grating'. These inclusion compounds are formed when small molecules, such as methanol, hydrogen sulphide or sulphur dioxide, are completely enclosed in cavities formed by a host compound, such as the β -quinol network (Fig. 1b). Here we have addition products with little or no direct attachment — and no covalent bonds — between the 'host' and the 'guest'. Powell's work was the beginning of what would

eventually become a major part of supramolecular chemistry — the design of host cages that allow the selective inclusion and expulsion of guest molecules. One of the oldest uses of clathrates is in crude oil refining, in which undesirable paraffins are removed from gasoline by trapping in clathrate lattices.

The early clathrates were discovered by chance, but rational design has led to enhanced properties. For example, a host matrix made from a copper-based polymer material absorbs and releases methane. This organic-inorganic hybrid competes with porous zeolites in its absorptive capacity, and could offer new applications for clathrates, such as the purification of drugs and trapping and storage of toxic materials⁵.

Chemists could not understand these inclusion compounds in terms of normal covalent bonding, and they were often relegated to the fringes of chemistry. But with the discovery of useful properties, chemists had to take these compounds seriously — the citadel of the isolated molecule was vulnerable after all.

Intermolecular interactions

The term supramolecular chemistry was coined in 1969 by Jean-Marie Lehn in his study of inclusion compounds and cryptands (Fig. 1c)². The award of the 1987 Nobel Prize in Chemistry to Charles Pedersen, Donald Cram and Lehn signified the formal arrival of the subject on the chemical scene. Lehn defined supramolecular

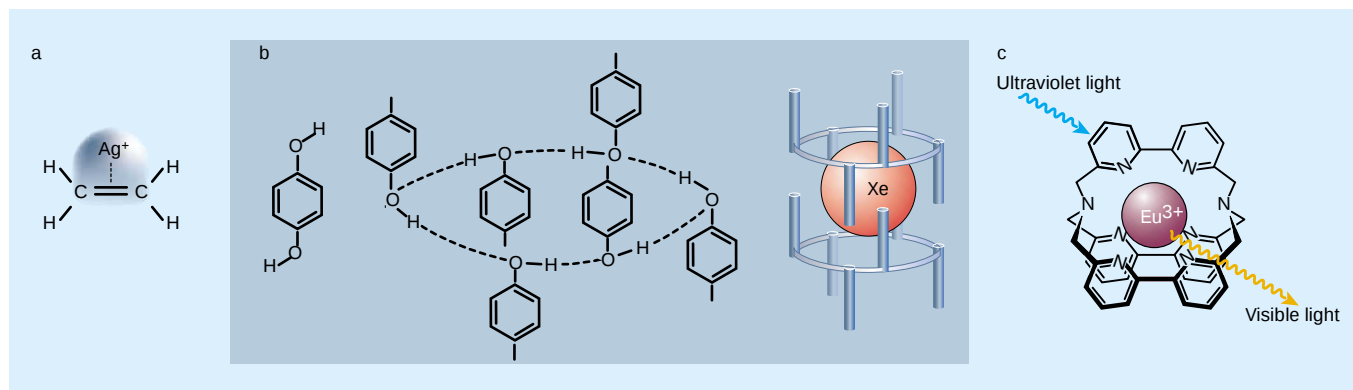
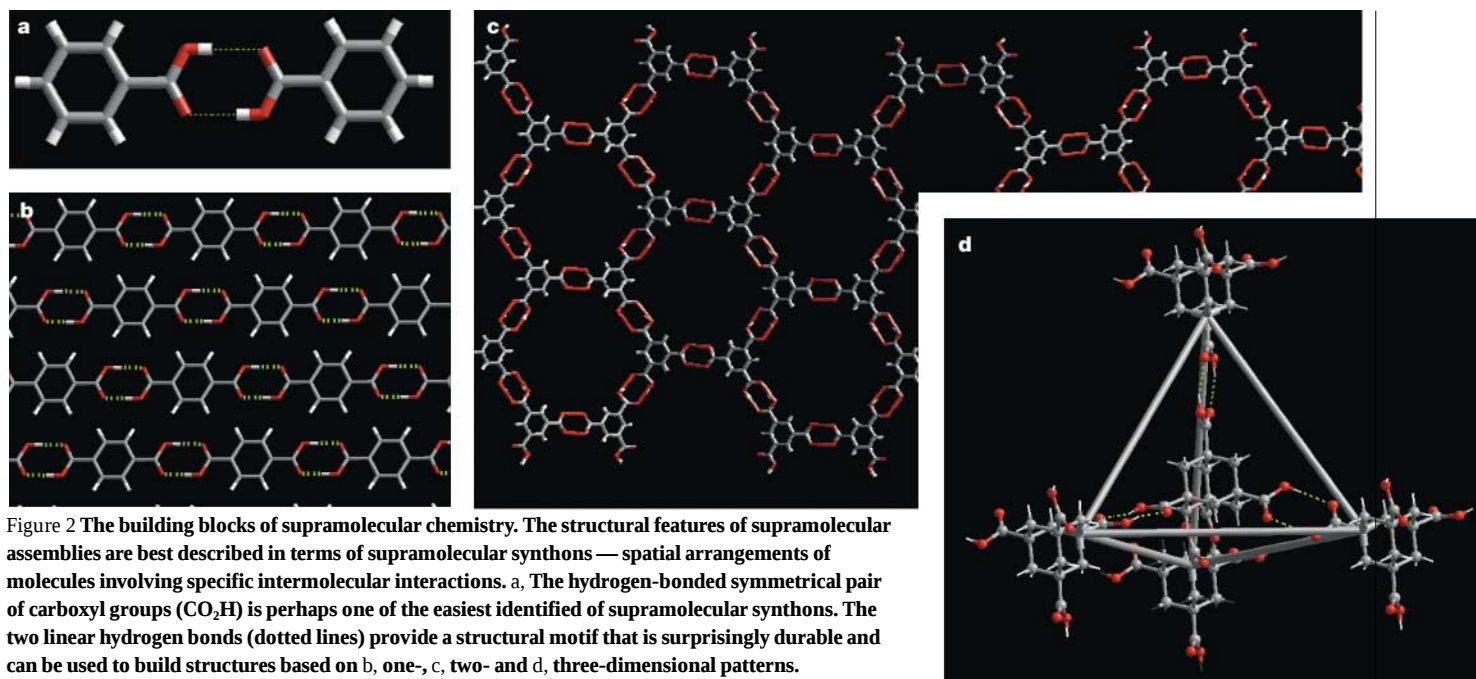


Figure 1 Supramolecular structures formed by intermolecular interactions. a, A donor-acceptor complex involving silver and ethene. b, Hydroquinone molecules assemble into a clathrate using hydrogen bonds. This means they can form solid-state host-guest complexes in which the hydroquinone network is the host and the guest is a small molecule, such as the xenon atom shown. c, A cryptand contains a spherical internal cavity studded with donor sites, suitable for enclosing a metal ion. Ultraviolet light absorbed by the cryptand shown here excites the metal ion, $\text{Eu}(\text{III})$, which then emits radiation at longer (visible) wavelengths.



chemistry as “the chemistry of the intermolecular bond”. Just as molecules are built by connecting atoms with covalent bonds, supramolecular compounds are built by linking molecules with intermolecular interactions.

Supramolecular structures are the result of not only additive but also cooperative interactions, and their properties generally follow from their supramolecular character (Boxes 1, 2). So even with the clathrates, their whole is more than the sum of their parts. These properties are important in both materials science (magnetism, conductivity, sensors, nonlinear optics) and biology (receptor–protein binding, drug design, protein folding).

In any supramolecular assembly, a large number of intermolecular interactions is possible — but only a few are actually observed. The weakness of these interactions makes it difficult to predict supramolecular structures and means that, in solution, supramolecular structures are not always stable over time. But this flexibility also means that they are frequently favoured in important mechanisms, notably in biological reactions and in crystallization processes, where the ability to form short-lived transition states and to perform trial-and-error correction easily is essential.

Intermolecular interactions are divided into two classes: isotropic, medium-range forces and anisotropic, long-range forces.

Isotropic forces define the shape of the individual molecules, as well as size and close packing of molecules, whereas anisotropic forces determine intermolecular orientations and functions. For example, the three-dimensional shapes of biomolecules, such as proteins and enzymes, are the result of medium-range intermolecular interactions. At a simple level, all molecular recognition can be said to arise from isotropic interactions, in other words by the fitting together of bumps and hollows among the components of the supramolecular structure. But most directional effects — and function is related to these effects — depend on the anisotropic interactions. Generally, the anisotropic interactions

Box 1 Hard applications

Supramolecular chemistry has always been associated with new materials and applications. Chemistry is driven by the desire for new functions, with the study of structure as a necessary first step towards the achievement of that goal. Ideally, useful materials would be designed by taking a single molecule and ‘sticking it’ to others of its kind to form three-dimensional assemblies. Implied in such a strategy is the ability to fine-tune function without necessarily disturbing structure. Accordingly, the total synthesis of a useful material can be dissected into molecular and supramolecular components. Traditional organic chemistry already provides all the necessary technology for the synthesis of the molecular

building blocks. Supramolecular synthesis, which requires manipulation of intermolecular interactions, is still evolving.

Most solid-state devices, such as electronics, require a degree of order that is only possible with crystalline materials. Unlike porous materials, crystals have densely packed molecules and any chemical change is likely to destroy the crystal and its properties. But organoplatinum molecules have successfully been engineered to make a crystalline material that reversibly binds sulphur dioxide (SO_2) gas¹⁸. When the colourless organoplatinum crystals are bathed in SO_2 they turn bright orange and their total volume increases by about a quarter, but the

crystals remain perfectly ordered. The SO_2 can be absorbed and expelled many times without loss of crystallinity. Part of the reason for this remarkable behaviour is that the crystalline framework is held together by supramolecular interactions — a string of hydrogen bonds — which can more easily tolerate such deformation. These organoplatinum crystals might find use as a gas-storage device, a sensor or even as an optical switch.

Nanocrystalline materials with ultrafine grains are potentially useful in molecular-scale electronics as magnetic¹⁹, semiconducting, dielectric and ferroelectric materials. Part of the problem chemists have with molecule-based systems is

connecting them to the macroscopic world. One approach is to modify a nanocrystalline surface with supramolecular components. For example, a ruthenium complex bound to nanocrystalline titanium dioxide gives a supramolecular structure that undergoes photochemically driven electron- and energy-transfer processes. This makes them attractive materials for use in solar cells²⁰. Supramolecular aggregates of organic molecules have also been used to make molecular switches, whose conductivity is controlled by the chemical state of the molecule. Chemical sensors based on such switches could in theory be used to detect individual molecules of a pollutant.

G.R.D.

involve partially charged atoms, such as nitrogen, oxygen, chlorine, phosphorus and sulphur.

Isotropic interactions include van der Waals forces, which act between all atoms and molecules. These can be repulsive or attractive depending on the distance between the interacting non-bonded atoms, and are responsible for gross supramolecular arrangements. Although these forces are individually weak — they have bond energies of 8 kJ mol^{-1} compared with the 400 kJ mol^{-1} of covalent bonds — they become significant when considered in numbers. This is the essence of supramolecular thinking. The remarkable ability of the house lizard to stick to a ceiling is a result of millions of van der Waals interactions formed by the spatulae at the ends of fine hairs that cover the soles of its feet⁶.

Although at a simple level molecular recognition can be said to hinge on isotropic interactions, at a higher level it is an anisotropic interaction that is the master key: the hydrogen bond⁷. In any hydrogen bond, $\text{X-H}\cdots\text{A}$, a hydrogen atom acts as a bridge between two atoms X and A. These atoms always tend to be negatively charged (electronegative), which gives the hydrogen bond an electrostatic character, as the electropositive hydrogen atom holds the negative atoms in thrall.

If X and A are both quite electronegative, for example in $\text{N-H}\cdots\text{O}$, the hydrogen bond is 'strong' or 'conventional' ($20\text{--}40 \text{ kJ mol}^{-1}$). But if either or both X and A are of moderate to weak electronegativity, such as in $\text{C-H}\cdots\text{O}$, the hydrogen bond is 'weak' or 'non-conventional' ($2\text{--}20 \text{ kJ mol}^{-1}$)⁸. In some systems, such as those involving the HF_2^- ion, the strength of the hydrogen bonds can reach quasi-covalent levels (170 kJ mol^{-1}). Hydrogen bonds can also form between more than two atoms. The importance of hydrogen bonds that are formed with double and triple bonds, such as $\text{C}=\text{C}$ and $\text{C}\equiv\text{C}$, is increasingly being recognized. For example, hydrogen bonds formed by groups such as OH, NH and CH with the double bonds in aromatic rings are recognized as being key in the stabilization of biomolecular structures⁹.

In general, the hydrogen bond is a composite interaction, which can have pronounced covalent, electrostatic or van der Waals components and consequently spans a wide energy range. The strength of interaction dictates the length and orientation of the hydrogen bond: short, linear bonds are almost always the strongest. But even weak bonds can be significant. Weak interactions tend to be hydrophobic, so they can persist in ionic solvents better than stronger hydrogen bonds.

Predicting supramolecular structures is hard, not only because of the sheer numbers of possible interactions involved, but also

because in energetic terms there is not much to choose between these various interactions. If one interaction is not much more energetically favourable than the others, then there is no clear winner to predict. The challenge for the supramolecular chemist attempting to synthesize these structures is to ensure that the molecules involved are oriented in such a way that maximizes the strength of the desired interactions. Two examples of instances where intermolecular interactions are crucial are crystal engineering and crystallization.

Crystal engineering

Organic crystals are supramolecular entities in that they are built from molecules. In crystal engineering, the supramolecular chemist aims to design and control packing arrangements to design crystals with specific properties. But this is not routinely possible from knowledge of the molecular structure alone. The best chemists can do is to find recurring packing patterns adopted by certain functional groups and rely on the robustness of such motifs to create new solid-state structures. These small repetitive units are called supramolecular synthons¹⁰. They economically but fully define the essence of a structure, and embody specific intermolecular interactions (Fig. 2). Repetition is what qualifies a structural motif to be considered a synthon and it is repetition that is the key to successful crystal engineering.

Even after identifying an important intermolecular interaction, failure to predict a crystal structure can arise because the same interaction can be used to make many different synthons. For example, the carboxyl group (CO_2H) usually assembles in crystals via pairs of $\text{O-H}\cdots\text{O}$ hydrogen bonds¹¹. But in cubane acids, the carboxyls form a synthon containing both $\text{O-H}\cdots\text{O}$ and $\text{C-H}\cdots\text{O}$ interactions (Fig. 3, overleaf).

Crystal engineering is shifting its focus from devising structures to designing properties. Property design can only follow from structure and our understanding of the principles by which molecules assemble into crystals is still rudimentary. Chemists are trying various computational and experimental approaches towards crystal design. And the synthon concept tries to simplify what is essentially a very difficult problem. But even with apparently straightforward synthons, things can go wrong.

In coordination polymers, for example, the relatively strong coordinate bonds (a type of covalent bond) between organic molecules and metal atoms are used to generate a solid network. This approach offers greater structural repetition because the coordinate bonds have both strength and explicit geometries. The first metal-organic networks were developed to mimic zeolites — microporous materials used for catalysis and separation processes. But even these

Box 2 Soft solutions

On the softer, non-crystalline side of supramolecular chemistry, desirable properties include solubility and chirality (molecular asymmetry). For example, a water-soluble polyfullerene has been synthesized using supramolecular methodology²¹. The reactants are prearranged in the cavity of a cyclodextrin, a container molecule, and then the polymerization is carried out. Potential biomedical applications of this supramolecular polymer follow from the fact that it scavenges the natural 'free radical', 1,1-diphenyl-2-picrylhydrazyl, more strongly than C_{60} itself, and that it cleaves DNA oligonucleotides in the presence of light.

Hydrogen-bonded systems often display chirality at the supramolecular level. Chemists want to control chirality so that supramolecular products are purely left- or right-handed isomers, and not a mixture of both. Supramolecular structures can now be designed to form sufficiently stable hydrogen bonds that pure-handed isomeric assemblies can be isolated²². The efficacy of drugs that are administered in crystalline form as fine powders can depend strongly on whether they have the correct isomeric structure and also the correct polymorphic structure.

Polymorphism is when a given molecule can exist in different crystalline forms, which are stable under different conditions. Polymorphs can therefore be thought of as supramolecular isomers, species in which the relative positioning of the same molecules is different. Polymorphs of a drug can have quite different properties. Their solubilities can be different, as well as their biological activity. In some cases, the less stable form will crystallize first, and then slowly transform, over time, into the more stable crystal. This can be a problem if the active form of a drug is the less stable polymorph, and turns into the more stable — but less active and perhaps even harmful — form during storage. In this context, can one design a molecule that is guaranteed to crystallize in a particular polymorphic form under physiological conditions? This is now a supramolecular challenge for the pharmaceutical industry.

G.R.D.

seemingly reliable synthons failed when the packing density was low — more often than not the porous structure was unstable to collapse. The latest generation of open networks, which use a dicarboxylate linker to help stabilize the structure, can survive the loss or exchange of guest molecules¹².

Crystallization of molecules in solution is also a supramolecular process. Like other chemical reactions, it is governed by thermodynamic and kinetic factors. In molecular chemistry, pure energy considerations — enthalpy — are usually the decisive factor, so chemists only have to worry about identifying the strongest and weakest bonds, because they usually want to make the strongest

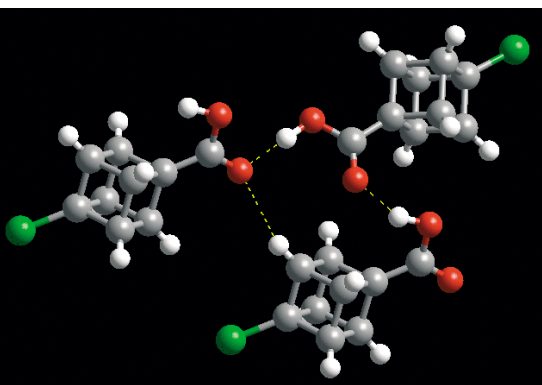


Figure 3 In many cubane acids, the carboxyl groups (CO_2H) form a synthon based on extended $\text{O}-\text{H}\cdots\text{O}$ and $\text{C}-\text{H}\cdots\text{O}$ bonds. In 4-bromocubanecarboxylic acid, for example, the $\text{C}-\text{H}\cdots\text{O}$ hydrogen bond provides the overall synthon with additional support¹¹. It is thought that without this interaction, the exotic topology could not be realized.

bonds and break the weakest. But in supramolecular chemistry, there are so many possible interactions that you have to consider both the enthalpy and entropy (disorder) inherent in every interaction. It is surprising, then, that although putative crystal structures can be similar in terms of energy, many organic molecules yield only one crystal structure under given conditions. Molecules seem to know how to crystallize into the correct structure, even if chemists cannot predict it.

Crystallization is a complex but highly efficient process in which a number of molecular groups compete with each other to be sites for intermolecular interactions that might lead to a stable crystal structure. Before the earliest stage of crystallization — known as nucleation — all molecular groups present in the solution try out different possible routes to alternative intermolecular interactions. But only a few of these interactions will be thermodynamically, kinetically or statistically sustained. The stable interactions lead to robust synthons and alternative possibilities are quickly and efficiently excluded. In this respect, there are similarities between the crystallization of a small organic molecule and the folding of a protein. Both proceed through intermediate, structure-determining clusters, resulting in a considerable increase in crystallization or folding efficiency.

A fascinating example of such a mechanism is provided by molybdenum chemistry. Molybdenum can form giant molecular clusters or rings with nanometre-sized cavities. Addition of two Mo_{36} 'hub-caps' to the wheel-like Mo_{176} cluster results in a Mo_{248} cluster¹³. The hub-caps are not stable in solution, showing that crystallization can proceed in steps, with complex structures forming from more simple ones that are not necessarily stable on their own.

Similarly, solvation — roughly the opposite to crystallization — is a phenomenon in which the enthalpic advantage of retaining solvent in a crystal outweighs the entropic advantage of expelling it into the bulk solvent¹⁴. An understanding of crystallization is important for the development of supramolecular chemistry and crystal engineering; both experimental and computational approaches (molecular dynamics) are currently being used to tackle this exciting problem.

A study of the protein apoferritin recently provided the first ever observations of protein nucleation¹⁵. By using atomic force microscopy, tiny clusters of apoferritin can be watched as they slowly grow or dissolve. Protein crystallization is more an art than a science, yet it is vital for determining protein structure.

A quiet revolution

Almost unobtrusively, chemistry is changing from a subject to a language that is needed to communicate core issues in biological and materials science. This change will surely lead to striking and unexpected practical applications, and supramolecular chemistry is at the vanguard (Boxes 1 and 2). Stephen Lippard recently made a 'wish list' for chemistry that represents this change¹⁶. Beyond what I have already discussed here, new goals for supramolecular chemistry include building porous structures — clathrates or coordination compounds — that have internal sites for catalysis, and designing solid-state reactions that are environmentally friendly because they are solvent free.

So far, the success stories include one

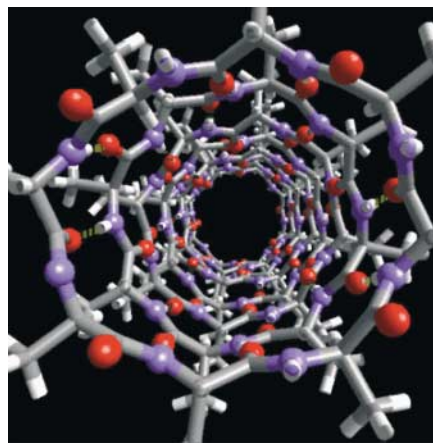


Figure 4 Supramolecular approach to a biological problem. Cyclic peptides with alternating D- and L-amino acids create a flat, ring-shaped synthon. When several cyclic peptides are stacked together they can self-assemble into a stable nanotube. The walls of the tube are β -pleated sheets. In mice, these nanotubes have selective antibacterial activity by increasing the permeability of bacterial membranes¹⁷. (The structure shown here features different side groups from those used by Fernandez-Lopez *et al.* in ref. 17.)

published on page 452 of this issue¹⁷. The authors of this paper have created supramolecular nanotubes that have important biological activity *in vivo* (Fig. 4). By synthesizing cyclic peptides consisting of alternating D- and L-amino acids, a synthon that self-assembles into a nanotube is created. The resulting nanotubes have selective antibacterial activity in mice by increasing the permeability of bacterial membranes. They are highly effective against drug-resistant bacteria, highlighting the advantages of a non-biological treatments over conventional ones.

Friedrich Wöhler's synthesis of urea in 1828, the first laboratory synthesis of a naturally occurring compound, symbolized the end of the vitalistic approach to chemistry — the idea that living organisms differ from non-living substances because they possess a 'vital force'. But with the arrival of Emil Fischer and supramolecular chemistry, chemists are now more than ever concerned with the transition from chemistry to biology. How do life processes work? The fantastic levels of specificity achieved by biological machines may be reduced to weak interactions, to chemical recognition and function, and inexorably down to physics itself. Yet, a reductionist approach is simplistic beyond the extreme. A scientifically more acceptable view of vitalism is that living and non-living matter differ not in content but rather in organizational complexity — and our understanding of this theme may well turn out to be the biggest breakthrough in supramolecular science. ■

Gautam R. Desiraju is at the School of Chemistry, University of Hyderabad, Hyderabad 500 046, India. e-mail: desiraju@uohyd.ernet.in

1. Fischer, E. *Ber. Dt. Chem. Ges.* **27**, 2985–2993 (1894).
2. Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives* (VCH, Weinheim, 1995).
3. Desiraju, G. R. *Nature* **408**, 407 (2000).
4. Powell, H. M. *J. Chem. Soc.* 61–73 (1973).
5. Noro, S., Kitagawa, S., Kondo, M. & Seki, K. *Angew. Chem. Int. Ed. Engl.* **39**, 2081–2084 (2000).
6. Autumn, K. *et al.* *Nature* **405**, 681–685 (2000).
7. Jeffrey, G. A. *An Introduction to Hydrogen Bonding* (Oxford Univ. Press, New York, 1997).
8. Desiraju, G. R. & Steiner, T. *The Weak Hydrogen Bond in Structural Chemistry and Biology* (Oxford Univ. Press, Oxford, 1999).
9. Steiner, T. & Koellner, G. *J. Mol. Biol.* **305**, 535–557 (2001).
10. Desiraju, G. R. *Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327 (1995).
11. Kuduva, S. S., Craig, D. C., Nangia, A. & Desiraju, G. R. *J. Am. Chem. Soc.* **121**, 1936–1944 (1999).
12. Li, H., Eddaoudi, M., O'Keefe, M. & Yaghi, O. M. *Nature* **402**, 276–279 (1999).
13. Müller, A., Shah, S. Q. N., Böge, H. & Schmidtman, M. *Nature* **397**, 48–50 (1999).
14. Nangia, A. & Desiraju, G. R. *Chem. Commun.* 605–606 (1999).
15. Yau, S.-T. & Vekilov, P. G. *Nature* **406**, 494–497 (2000).
16. Lippard, S. J. *Chem. Eng. News* 7 August 64–65 (2000).
17. Fernandez-Lopez, S. *et al.* *Nature* **412**, 452–455 (2001).
18. Albrecht, M., Lutz, M., Spek, A. L. & van Koten, G. *Nature* **406**, 970–974 (2000).
19. Kahn, O. *Acc. Chem. Res.* **33**, 647–657 (2000).
20. Freemantle, M. *Chem. Eng. News* 19 June, 48–58 (2000).
21. Samal, S., Choi, B.-J. & Geckeler, K. E. *Chem. Commun.* 1373–1374 (2000).
22. Prins, L. J., De Jong, F., Timmerman, P. & Reinhoudt, D. N. *Nature* **408**, 181–184 (2000).

Persistence of visual memory for scenes

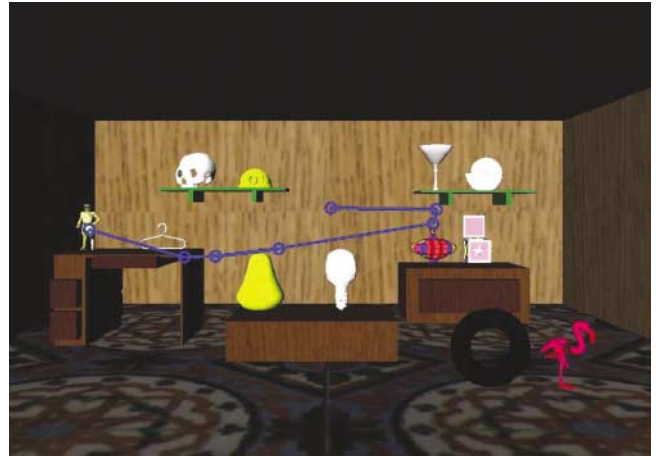
A medium-term memory may help us to keep track of objects during visual tasks.

Building a complete representation of a visual scene requires information to be remembered across separate glances and over time^{1,2}, but it has been suggested that visual details are forgotten soon after they are viewed^{3–6}. Here I show that cumulative memory build-up allows the same number of objects to be recalled, irrespective of whether these were seen in a series of short, separate presentations several minutes apart, or as one continuous presentation of the same total duration. I find that the build-up of visual memory over time is much better than has been widely thought and may underlie the successful performance of real-world visual and cognitive tasks that require people to keep track of objects in the immediate environment.

Scenes were computer-generated 'rooms' containing 12 unrelated objects (Fig. 1). They were presented for a cumulative viewing duration of 1, 2, 3 or 4 seconds, either all at once (continuous-presentation trials) or as brief views of 0.25, 1 or 2 seconds in duration and separated by as many as eight other trials (re-test trials). Participants ($n=6$; vision was normal or corrected to normal) did not know whether, or when, a particular scene would be re-tested, or whether a trial would contain an old or a new scene. The long intervals between re-tests (0.5–4.0 min), the presence of intervening presentations, and the random occurrence of re-tests precluded effective rehearsal.

Memory improved with re-testing. The number of items recalled after a series of separate brief views was almost identical to the number recalled after a single continuous

Figure 1 Visual memory for objects in a scene. Representative computer-generated display of 12 objects ($1-2^\circ$ of visual angle) in a room ($10 \times 8^\circ$ of visual angle). The blue line shows the scanning path followed by one subject. Objects were presented for a cumulative viewing duration either as a continuous presentation or as brief separate views summing to the same duration. The number of items recalled was similar in both cases.



view of the same total duration (Fig. 2). For example, the number of items recalled after the last of four separate 1-second trials was equal to the number of items recalled after 4 seconds of continuous presentation. Memory of each scene continued to accumulate over repeated viewings as though the scene had never been out of sight.

I investigated the nature of memory accumulation by presenting new objects on previously viewed backgrounds. If the entire scene (objects plus background) is encoded in the memory, then a familiar background should bring to mind the memory of the original set of objects and interfere with memory for the new set of objects⁷, which I found to be the case.

Memory was poorer for new objects seen against old backgrounds than for completely new scenes ($P < 0.05$; 6 subjects). When object names, such as 'apple', were presented in place of a picture of the object, the memory accumulation and the effect of repeating previously viewed backgrounds were both virtually abolished. These results suggest that a visuo-spatial representation of the whole scene was remembered across re-tests, rather than simply a verbal list of object names.

Earlier studies of visual memory produced diverse estimates of capacity^{4–6}, with memory being either inferred from performance of a concurrent task or assessed using displays containing semantic cues that may have influenced encoding^{8–10} or guessing¹¹ strategies. The experiments I describe here remove semantic cues and evaluate memory capacity by using measures of recall rather than by a secondary task.

The visual memory reported here is unusual in that it does not resemble traditional short- or long-term memory. Short-term memory is not involved because the time between re-tests of the same display

exceeded the temporal limits of information summation¹² and short-term memory¹³. Also, the presence of intervening scenes during the intervals between re-tests precluded rehearsal as a means of retaining the items in short-term storage. Traditional long-term memory was not involved either, because there was no build-up across days. The memory studied here may therefore be best described as 'medium-term' or 'disposable'.

Medium-term memory may underlie the ability to keep in mind the identity and location of objects while performing visuo-motor tasks that last for a few minutes within an unchanging visual environment^{2,14,15}. A medium-term visual memory could be instrumental in quickly directing the eye or arm to selected objects without the continual need for expensive or time-consuming visual searching.

David Melcher

Department of Psychology, Rutgers University,
Piscataway, New Jersey 08854, USA
e-mail: melcher@ruccs.rutgers.edu

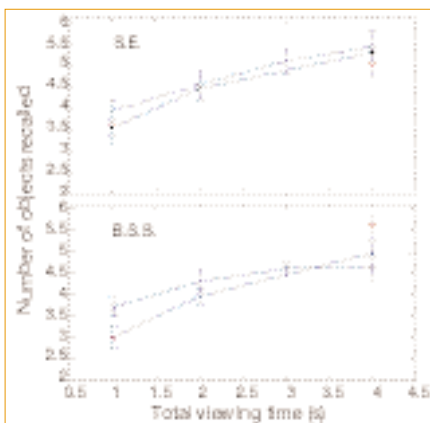


Figure 2 Number of items recalled as a function of total viewing time by two representative participants (S.E. and B.S.S.). Solid lines, continuous trials of 1, 2 or 4 s; dashed lines, performance after 1, 2, 3 or 4 re-test trials of 1 s each; squares, 4 trials of 250 ms each; triangles, 2 trials of 2 s each. Error bars represent standard error. Four other participants generated similar results (data not shown).

1. Dennett, D. C. & Kinsbourne, M. *Behav. Brain Sci.* **15**, 183–247 (1992).
2. Land, M. & Furneaux, S. *Phil. Trans. R. Soc. Lond. B* **352**, 1231–1239 (1997).
3. O'Regan, J. K. *Can. J. Psychol.* **46**, 461–488 (1992).
4. Ballard, D. H., Hayhoe, M. M. & Pelz, I. B. *J. Cogn. Neurosci.* **7**, 66–80 (1995).
5. Horowitz, T. S. & Wolfe, J. M. *Nature* **394**, 575–577 (1998).
6. O'Regan, J. K., Rensink, R. A. & Clark, J. J. *Nature* **398**, 34 (1999).
7. Chun, M. M. & Nakayama, K. *Vis. Cogn.* **7**, 65–81 (2000).
8. Friedman, A. J. *Exp. Psychol. Gen.* **108**, 316–355 (1979).
9. Pashler, H. *Percept. Psychophys.* **44**, 369–378 (1988).
10. Hollingsworth, A. & Henderson, J. M. *Vis. Cogn.* **7**, 213–235 (2000).
11. Miller, M. B. & Gazzaniga, M. S. *Neuropsychologia* **36**, 513–520 (1998).
12. Loftus, G. R., Duncan, J. & Gehrig, P. J. *Exp. Psychol. Hum. Percept. Perform.* **18**, 530–549 (1992).
13. Baddeley, A. in *The Oxford Handbook of Memory* (eds Tulving, E. & Craik, F. I. M.) 77–92 (Oxford Univ. Press, New York, 2000).
14. Hayhoe, M., Bensinger, D. G. & Ballard, D. H. *Vis. Res.* **38**, 215–237 (1998).
15. Epelboim, J. et al. *Vis. Res.* **35**, 3401–3422 (1995).

Bone histology

Evolution of growth pattern in birds

Living (neornithine) birds grow up rapidly and without interruption, terminating their growth within one year and, with a few secondary exceptions, starting to fly only after or near the completion of growth. Bone histology has revealed that pre-avian theropods also grew fast for most of the postnatal period, but that this growth was usually intermittent and probably extended for more than one year^{1–3}. We have found surprising evidence for an early postnatal slowing-down of growth in two lineages of flying basal birds, which suggests that birds may have started their evolution as precocious fliers.

The living birds and their Mesozoic relatives⁴ (all classified as Euornithes or as redefined Ornithurae) typically have highly vascularized, rapidly formed fibrolamellar bone, which is encircled peripherally (as well as endosteally) by a narrow layer of slowly formed lamellar bone (Fig. 1d). Fibrolamellar bone also predominates in the long bones of non-avian theropods^{1–3}. In contrast, our study shows that *Rahonavis*, a late Malagasy survivor of basal birds, and

the Enantiornithes, the dominant group of Cretaceous land birds, differ in their long-bone histology from both non-avian maniraptorans and euornithine birds in having much less or no fibrolamellar bone.

In the femur of *Rahonavis*, the outer layer of slowly formed lamellar bone is very thick relative to that of euornithine birds and makes up roughly 44% of the bone wall (Fig. 1c). The femora of adult enantiornithines (Fig. 1a) are exclusively composed of a lamellar type of bone tissue^{1,5}. However, the tibia of a perinatal enantiornithine specimen assigned to *Gobipteryx* (ZPAL MgR –/90; ref. 6) shows only fibrolamellar tissue with developing primary osteons (Fig. 1b), which is consistent with external evidence of fast-growing bone in an Enantiornithine hatchling⁷. We conclude that the fibrolamellar bone of enantiornithine hatchlings must have been completely removed by resorption and remodelling during their slow post-embryonic growth, and that the slowing down of growth must have occurred earlier in their ontogeny than in that of euornithines and possibly of all other birds.

The long bones of non-avian theropods usually show lines of arrested growth (LAGs), which mark interruptions of bone deposition^{1–3}. The average growth rate during their entire postnatal development was

therefore lower than those in euornithine birds. LAGs are also present in the enantiornithines (Fig. 1a) and the lamellar layer of *Rahonavis* (Fig. 1c), which indicates that the postnatal growth rate of these birds was even lower than the deposition rate of lamellar bone within a single growth spell.

The growth of young birds is known to be slowed down by locomotor activity, which probably sequesters substantial energy resources and accelerates the maturation of tissues⁸, especially bone⁹. Accordingly, among living land birds other than large ratites, growth rates are lowest in the most precocious fliers: that is, galliforms (particularly superprecocial megapodes) and tinamous^{10,11}. An early change in slow-growing bone tissue therefore provides new evidence of the superprecocial onset of flight in the Enantiornithes, which has been inferred from the relative length and almost complete ossification of the wing skeleton (including the pectoral girdle) in late embryos or hatchlings assigned to *Gobipteryx*⁶.

The principal and primary function of the wings in Enantiornithes and other flying birds is flight. The degree of ossification and the adult proportions of the wing skeleton in the hatchlings assigned to *Gobipteryx* therefore unequivocally indicates an early onset of flight even if some of the living birds re-evolved precocious flight by other means¹². On the basis of the remarkably consistent evidence from pelvic-limb histology and pectoral morphology, we suggest that the marked slow-down of growth in Enantiornithes was caused by the early onset of flight.

The phylogenetically earliest record of birds in which fibrolamellar bone is deposited over the entire growth period comes from *Patagopteryx*^{1,5}, a flightless, chicken-sized bird from the Lower Cretaceous period in South America. However, fibrolamellar bone of *Patagopteryx* continues to be deposited after a LAG, which may mark the ancestral separation between the phases of fast and slow growth. The growth pattern of *Patagopteryx*, a close relative of euornithines¹³, may therefore have been intermediate between basal and euornithine birds. Moreover, if precocious flight was responsible for reducing postnatal growth rate in early avian evolution, the loss of flight in primitive birds may have facilitated a reversion to the high, pre-avian growth rate. Thus, the flightlessness of *Patagopteryx* places a caveat on phylogenetic interpretation of its bone histology.

As the monophyly of *Rahonavis* and Enantiornithes can safely be ruled out, our study suggests that birds evolved through a stage of slowed postnatal growth and precocious flight, and that fast-growing chicks with underdeveloped flight apparatus seem to be close to the ancestry of euornithines.

Anusuya Chinsamy*, Andrzej Elzanowski†

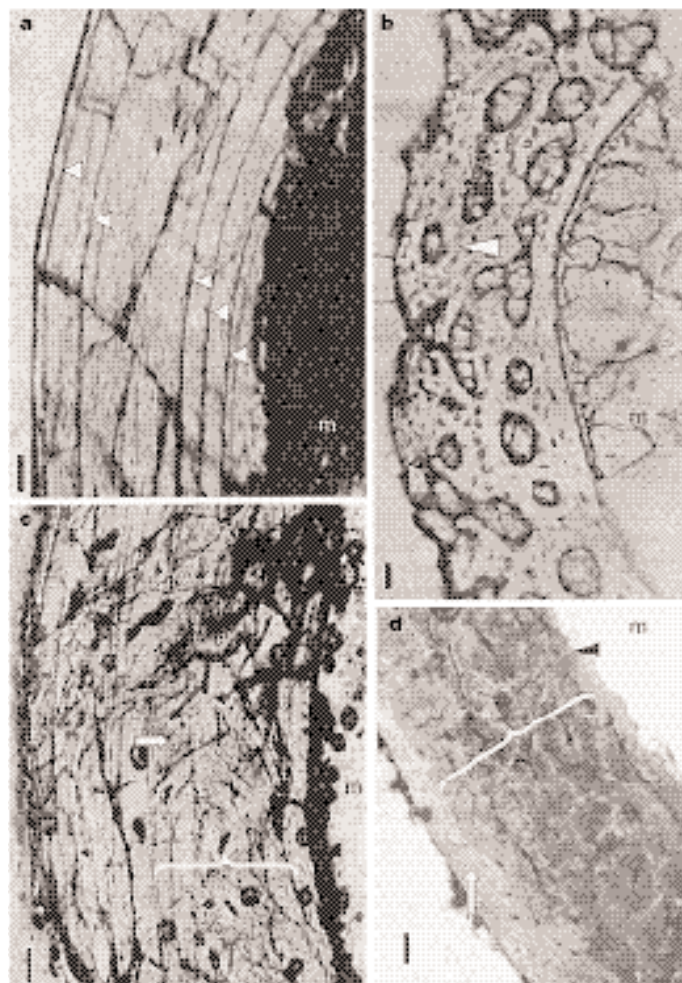


Figure 1 Bone-histology evidence (transverse sections). **a**, Femur of an Enantiornithine from the Maastrichtian Lecho formation, Argentina, showing poorly vascularized bone tissue with parallel fibres. Arrowheads indicate lines of arrested growth (LAGs). **b**, Tibia of a late embryonic skeleton of *Gobipteryx*⁷; arrowhead indicates woven bone matrix around channels that will later be filled by centripetal bone deposition. **c**, Femur of *Rahonavis* showing a narrow layer of endosteal lamellar bone, a region of fibrolamellar bone (bracket) and a double LAG (arrow), which marks the onset of deposition of the thick outer layer of parallel-fibred bone. **d**, Femur of *Diomedea chrysostoma* showing a layer of endosteal lamellar bone (arrowhead), a region of fibrolamellar bone (bracket) and a narrow outer layer of lamellar bone (arrow). m, medullary cavity. Scale bars: **a**, 47 μ m; **b**, 50 μ m; **c**, 80 μ m; **d**, 100 μ m.

*Department of Zoology, University of Cape Town, Private Bag, Rondebosch 7700, South Africa, and South African Museum, PO Box 61, Cape Town 8000, South Africa

†Institute of Zoology, University of Wrocław, Ulica Sienkiewicza 21, 50335 Wrocław, Poland
e-mail: elzanowski@biol.uni.wroc.pl

- Chinsamy, A. *Palaeont. Afr.* **27**, 77–82 (1990).
- Reid, R. E. H. *Mod. Geol.* **18**, 391–406 (1993).
- Varricchio, D. J. *J. Vert. Paleont.* **13**, 99–104 (1993).
- Houde, P. *Auk* **104**, 125–129 (1987).
- Chinsamy, A., Chiappe, L. M. & Dodson, P. *Nature* **368**, 196–197 (1994).
- Elzanowski, A. *Palaeont. Pol.* **42**, 147–179 (1981).
- Sanz, J. L. *et al.* *Nature* **276**, 1543–1546 (1997).
- Ricklefs, R. E. in *Avian Biology* Vol. 7 (eds Farner, D. S., King, J. R. & Parkes, K. C.) 1–83 (Academic, New York, 1983).
- Starck, J. M. *J. Morph.* **222**, 113–131 (1994).
- Ricklefs, R. E. *Ibis* **115**, 177–201 (1973).
- Starck, J. M. & Ricklefs, R. E. (eds) *Avian Growth and Development* 247–265 (Oxford Univ. Press, New York, 1998).
- Starck, J. M. & Ricklefs, R. E. (eds) *Avian Growth and Development* 3–30 (Oxford Univ. Press, New York, 1998).
- Norell, M. A. & Clarke, J. A. *Nature* **409**, 181–184 (2001).

Seed dispersal

Directed deterrence by capsaicin in chillies

The primary function of ripe, fleshy fruit is to facilitate seed dispersal by attracting consumers^{1,2}, yet many fruits contain unpleasant-tasting chemicals that deter consumption by vertebrates³. Here we investigate this paradox in the chilli (*Capsicum*) and find that capsaicin, the chemical responsible for the fruit's peppery heat⁴, selectively discourages vertebrate predators without deterring more effective seed dispersers.

The worldwide popularity of chillies has prompted numerous investigations of the biological properties of capsaicin^{5,6}, yet its evolutionary significance in the plants that produce it has not been examined. Chillies are low shrubs with fruit that is accessible to mammals as well as birds. However, capsaicin in the fruit has been found to repel or poison mammals, but not birds, in laboratory studies^{7,8}. This taxonomic distinction is predicted by the 'directed deterrence' (toxicity) hypothesis, whereby chemicals in ripe fruit function selectively to discourage seed predators without deterring beneficial seed dispersers^{9,10}.

To test the value of capsaicin as a directed deterrent, we used a native population of chiltepine chillies (*Capsicum annuum* var. *glabrusculum*) in southern Arizona¹¹ and determined which of the local animal species consumed chillies and which were put off by capsaicin. We then investigated the fitness consequences to the plant of seed consumption by both types of potential consumer.

Video observation of chilli plants (146 h in total) revealed that only birds fed on chillies diurnally, with curve-billed thrashers (*Toxostoma curvirostre*) being responsible for 72% of all fruit removal. To assess

nocturnal consumption, we presented fruits of both chilli and desert hackberry (*Celtis pallida*, the most common fruiting plant at our study sites) on microsites available to birds and mammals. During the day, both fruit types were removed in equal amounts (paired- $t_{26}=0.71$, $P=0.483$); at night, only hackberry fruits were taken (paired- $t_{20}=3.11$, $P=0.006$). Wild chillies thus seem to be consumed exclusively by birds and avoided by small mammals.

To determine whether small mammals avoid chillies because of their capsaicin content, we captured cactus mice (*Peromyscus eremicus*) and packrats (*Neotoma lepida*), the most abundant fruit-consuming mammals, and curve-billed thrashers from the chilli site and studied them in laboratory feeding trials. Animals received three fruit types simultaneously: *Capsicum annuum* from the study site; a non-pungent mutant variety of *Capsicum chacoense*, a chilli that is similar in size, colour and nutritive content to *C. annuum* but lacking in capsaicin as the result of a developmental mutation; and hackberry fruit as a control. Both mammal species consumed all hackberry, an intermediate amount of non-pungent chilli, and no pungent chilli; in contrast, thrashers consumed virtually all fruits and were not deterred by the presence of capsaicin (Fig. 1a). Thus, capsaicin selectively influences the feeding preferences of consumers.

Under the directed-deterrence hypothesis, deterred species (mammals) should be seed predators or ineffective seed dispersers, whereas undeterred consumers (thrashers) should increase the fitness of the chilli plant. We therefore examined the effects of fruit ingestion on seed germination by using non-pungent *C. chacoense*, as mammals will not eat hot chillies. Consumption of *C. chacoense* by both mammal species resulted in zero germination, whereas consumption by thrashers resulted in germination rates similar to those of control seeds planted directly from fruit (Fig. 1b). When pungent chillies were fed to thrashers, these seeds also germinated at the same rate as control seeds ($F_{1,10}=0.233$, $P=0.64$). These findings indicate that fruit-consuming small mammals are seed predators, whereas thrashers are not.

The benefit to the plant of fruit consumption by thrashers depends on whether thrashers deposit seeds in places that are suitable for germination, establishment, fruit maturation and dispersal. Although less than 40% of our study area was shaded, 86% of adult chilli plants ($n=309$) and all seedlings ($n=22$) were found under the shade of other shrubs, as a result of directed dispersal by thrashers and increased seedling survival in shaded areas¹¹.

However, not all shade is equal: shrubs and trees producing bird-dispersed fruits contributed less than 5% of all shade at our

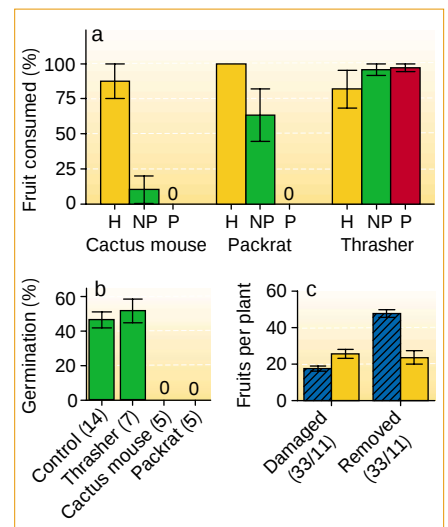


Figure 1 The effects of capsaicin on consumers and of consumers on chillies. **a**, In captive feeding-preference trials, the two most common mammals (five animals per species) readily consumed hackberry fruit (H); consumed non-pungent chillies (NP) only if they had not first sampled pungent chillies; and always avoided pungent chillies (P) (cactus mouse, $F_{2,12}=8.05$, $P=0.008$; packrat, $F_{2,12}=22.29$, $P<0.0005$). Thrashers (ten birds) ate virtually all pungent chillies that they were offered ($F_{2,15}=1.68$, $P=0.219$) and showed no aversion to capsaicin. **b**, Non-pungent chilli seeds from fruit ingested by thrashers germinated at the same rate as controls, whereas seeds from fruit ingested by cactus mice or packrats did not germinate, primarily because both mammal species rarely passed whole seeds ($F_{2,25}=19.0$; thrashers vs control $P=0.593$, mammals vs control $P<0.0005$). The number of animals used is shown in parentheses. **c**, Chilli plants growing under shrubs with bird-dispersed fruit (striped bars) had fewer damaged fruits ($F_{1,41}=8.2$, $P=0.006$) and more fruit removed ($F_{1,41}=32$, $P<0.0005$) than chillies growing in the shade of plants that do not produce bird-dispersed fruit (orange bars), after controlling for total fruit production. The respective numbers of plants surveyed are shown in parentheses.

study sites, yet 53% of chilli plants were growing under these plants ($\chi^2=24.9$, $P<0.0005$). Chillies under these plants gain two fitness advantages. First, they suffer significantly less insect damage to their immature fruit than do chillies that are shaded by other species (Fig. 1c). Damaged fruits are not released from the petiole and so the seeds are not dispersed; also, only 11% of seeds from damaged fruit germinate, compared with 70% of seeds from healthy fruit (paired- $t_{1,9}=4.4$, $P=0.002$). Second, because thrashers primarily forage in the canopies of fruiting trees and shrubs, removal and dispersal of healthy fruit from chillies growing beneath these species is higher than that from chillies under non-bird-dispersed plants (Fig. 1c). In short, directed dispersal by thrashers to sites under bird-dispersed plants is highly beneficial to chillies.

The capsaicin-mediated dispersal of chillies described here provides the first evidence, to our knowledge, that directed deterrence in a ripe fruit may shape plant-vertebrate interactions. The combination of chemical-mediated 'filtering' of

vertebrate seed predators¹⁰, together with directed dispersal by thrashers, creates a pattern of plant recruitment that is beneficial to wild chilli plants: mammalian seed-predation is eliminated and seeds are disproportionately dispersed to specific shaded areas where seedlings can be more easily established, predispersal fruit predation is lower, and seed dispersal is greater.

Joshua J. Tewksbury*, **Gary P. Nabhan†**

**Department of Biological Sciences, University of Montana, Missoula, Montana 59812, USA*

Present address: Department of Zoology, Box 118525, University of Florida, Gainesville, Florida 32611-8525, USA

e-mail: jtewksbury@zoo.ufl.edu

†Center for Sustainable Environments, Northern Arizona University, PO Box 5765, Flagstaff, Arizona 86011-5765, USA

1. Snow, D. W. *Ibis* **113**, 194–202 (1971).
2. McKey, D. in *Coevolution of Animals and Plants* (eds Gilbert, L. E. & Raven, P. H.) 158–191 (Univ. Texas Press, Austin, 1975).
3. Herrera, C. M. *Am. Nat.* **120**, 218–241 (1982).
4. Cordell, G. A. & Araujo, O. E. *Ann. Pharmacother.* **27**, 330–336 (1993).
5. Govindarajan, V. S. & Sathyanarayana, N. M. *Crit. Rev. Food Sci. Nutr.* **29**, 435–474 (1991).
6. Caterina, M. J. et al. *Nature* **389**, 816–824 (1997).
7. Mason, J. R., Bean, N. J., Shah, P. S. & Clark, L. J. *Chem. Ecol.* **17**, 2539–2551 (1991).
8. Norman, D. M., Mason, J. R. & Clark, L. *Wils. Bull.* **104**, 549–551 (1992).
9. Jansen, D. H. *Evolution* **23**, 1–27 (1969).
10. Cipollini, M. L. & Levey, D. J. *Am. Nat.* **150**, 346–372 (1997).
11. Tewksbury, J. J. et al. *Conserv. Biol.* **13**, 98–107 (1999).

Nanostructure

Epitaxial diamond polytypes on silicon

Carbon is unique in the variety of configurations it can adopt with itself and other elements. Here we show how ion beams can be used to nanostructure various diamond polytypes, epitaxially aligning them to a silicon substrate. The ready controllability of ion beams, which are already used to manufacture submicrometre-scale devices, means that our findings should enable new carbon and non-carbon materials to be nanostructured for a host of applications.

Metastable phases (including carbon) can be deposited using hyperthermal species (10–100 eV) by a ‘shallow-implantation’ process¹. The energy barrier for their formation is overcome by trapping energetic species in subsurface positions of a ‘mould’ host matrix. Examples include amorphous, diamond-like carbon films with a local tetrahedral configuration²; transformation of carbon nanotubes to cubic diamond by electron bombardment³; and ion-beam nanostructuring of multi-walled, small-diameter carbon nanotubes⁴.

Diamond polytypes are formed by changing the stacking sequence of the closed-packed plane of the diamond crystal. We used ion beams to form diamond poly-

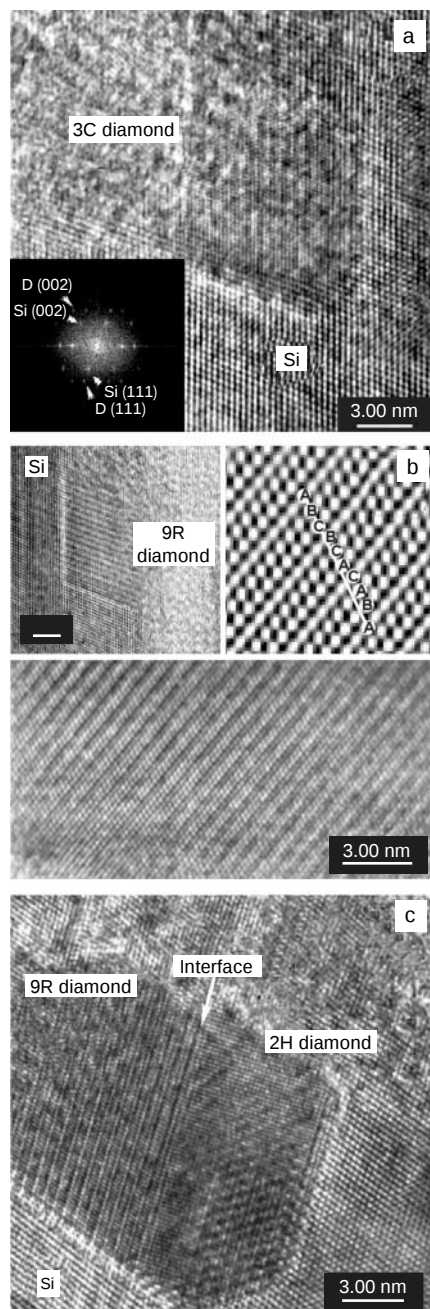


Figure 1 High-resolution transmission electron microscopy (Philips CM200) cross-sectional images of diamond polytypes. **a**, Cubic-diamond crystallite grown epitaxially on a silicon step; inset, crystallite diffractogram (Fourier transform). There is no misorientation between the diamond and silicon. **b**, Top left, 6-nm 9R diamond crystallite grown epitaxially on a Si step; bottom, part of a 60-nm 9R crystallite; top right, filtered image of 9R diamond showing the 9R repeat ‘ABCBCACAB’ sequence. **c**, Another diamond crystallite, also grown epitaxially on a Si step. Note the different symmetry of the 2H crystallite (right) and the sharp boundary with the 9R crystallite (left).

types (cubic, 3C; hexagonal, 2H; and rhombohedral, 9R, where the number denotes the number of closed-packed planes in a unit cell⁵) differing in the stacking order and periodicity of their closed-packed planes. The stacking sequence is controlled by the mould provided by the nucleation site^{1,6} and by the deposition conditions.

Films were prepared using a (CH₄ or C₂H₂)/Ar/H₂-fed Kaufmann source to bombard a Si(100) sample held at 600 or 700 °C with ions at 80 or 200 eV. The Fourier transform diffractogram of a cubic diamond crystal nucleated on a silicon step indicates the high-quality cubic-to-cubic epitaxial relationship between the Si and diamond, with a (111) diamond/Si interface and the absence of misorientation (Fig. 1a).

Figure 1b shows an epitaxial crystal and part of a similar large crystal, with its filtered image: the 9R registry is evident. Orientation relations for this phase are: [0001]_{9RDia}//[-111]_{Si}, [10-10]_{9RDia}//[[110]_{Si}, [01-10]_{9RDia}//[0-11]_{Si} (hexagonal indexing).

The left part of a third crystallite (Fig. 1c) is 9R diamond, but its right part has a different stacking sequence and symmetry from 9R and cubic diamond (Fig. 1a, b). The lower-right portion of this crystallite gives a complex image (Moiré fringes, resulting from interference between beams diffracted from both the silicon and the diamond lattices). The upper-right part shows a grid of lines normal to each other, in contrast to the 109.5° angle seen in the high-resolution transmission electron micrographs of cubic diamond (Fig. 1a) and silicon (Fig. 1a, c). Comparison with image simulations for possible diamond polytypes indicates that this crystallite is a hexagonal polytype; the nucleation on a silicon step and the strong epitaxial relations (the same as those of the 9R polytype) are evident. Although these crystallites may contain defects, we have shown that different diamond polytypes can be structured, sometimes by sharp transition from one polytype to another (Fig. 1c).

Polytypes have similar properties with variations that may be useful⁷. Heterojunctions are an example of an application that can involve different polytypes⁷. Our findings indicate that metastable phases can be manipulated by using low-energy ion beams, offering new opportunities to control material synthesis.

Y. Lifshitz*, **X. F. Duan***, **N. G. Shang***, **Q. Li***, **L. Wan†**, **I. Bello***, **S. T. Lee***

**Center of Super-Diamond and Advanced Films, and §Department of Physics and Materials Science, City University of Hong Kong, China*
e-mail: apannale@cityu.edu.hk

†Beijing Laboratory of Electron Microscopy, Center for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Hong Kong SAR, China

†On leave from: Soreq Nuclear Research Center, Yavne 81800, Israel

1. Lifshitz, Y., Kasi, S. R. & Rabalais, J. W. *Phys. Rev. Lett.* **62**, 1290–1293 (1989).
2. Lifshitz, Y. *Diamond Rel. Mater.* **8**, 1659–1676 (1999).
3. Banhart, F. & Ajayan, P. M. *Nature* **382**, 433–435 (1996).
4. Peng, H. Y. et al. *Appl. Phys. Lett.* **77**, 2831–2833 (2000).
5. Ramsdell, R. S. *Am. Mineral.* **32**, 64–82 (1947).
6. Lee, S. T. et al. *Science* **287**, 104–106 (2000).
7. Morkoc, H. et al. *J. Appl. Phys.* **76**, 1363–1398 (1994).

Dinosaurian growth rates and bird origins

Kevin Padian*, Armand J. de Ricqlès† & John R. Horner‡

* Department of Integrative Biology and Museum of Paleontology, University of California, Berkeley, California 94720-3140, USA

† Équipe Formations Squelettiques, UMR 8570 CNRS, Université Paris VII, 75251 Paris cedex 05, France, and Collège de France, Paris, France

‡ Museum of the Rockies, Montana State University, Bozeman, Montana 59717-0040, USA

Dinosaurs, like other tetrapods, grew more quickly just after hatching than later in life. However, they did not grow like most other non-avian reptiles, which grow slowly and gradually through life. Rather, microscopic analyses of the long-bone tissues show that dinosaurs grew to their adult size relatively quickly, much as large birds and mammals do today. The first birds reduced their adult body size by shortening the phase of rapid growth common to their larger theropod dinosaur relatives. These changes in timing were primarily related not to physiological differences but to differences in growth strategy.

For most of the 160 years since dinosaurs were named, they have been considered typical, if overgrown, reptiles, and it was assumed that they grew at rates broadly similar to those of extant reptiles¹. Indeed, there was no particular reason to think otherwise, even though their upright stance and parasagittal gait led Richard Owen, in naming them, to put them in a completely new group of Reptilia. However, in the 1970s interest in the question of dinosaur metabolism renewed ideas about their growth strategies and rates^{2–5}. Some physiological models concluded that dinosaurs could have been perfectly ‘good reptiles’⁶ whose large size gave them the metabolic benefits of inertial homoiothermy without truly departing from an ectothermic strategy⁷. However, a growing body of literature on bone growth rates suggests that, whereas dinosaurs may not have grown at rates exactly like those of extant birds and mammals, they seem generally to have been more like them than like other extant reptiles.

Evidence for this shift in perspective comes largely from bone histology, the microstructural characteristics of bone that reflect ontogenetic, environmental, mechanical and phylogenetic factors⁸. Recently renewed efforts have begun to use these routinely sampled bone tissues to calculate growth rates^{9–20}.

Timing the growth of extinct animals

To time the growth of extinct animals, two lines of evidence are generally used, and both depend on homologous features in extant vertebrates and on simple actualistic reasoning (using the features of extant organisms to interpret the fossil record). First, the type of primary (appositional) bone tissue indicates the overall range of growth rate. A given type of primary bone tissue grows at the same range of rates in any taxon²¹; hence, it is presumed that a certain kind of tissue deposited in an extinct animal would have grown at the approximate rates that it does in animals today. Second, bone growth is often punctuated by the deposition of various rings, lines or annular structures that may reflect cyclical annual growth, seasonal stress or other endogenous rhythms^{2,4,22,23}, so they may be skeletochronological indicators. However, consideration of extant vertebrates demonstrates that the biological meaning of these indicators must be assessed on a case-by-case basis. Some growth rings are of relatively avascular bone that interrupts vascular tissues, representing a real change in the type of deposition^{2,4}; others merely comprise lines of temporary cessation of growth and even slight erosion of the bone surface^{2,11,24}; and some simply represent indefinite pauses with no erosion^{9,16,24}. Generally, ectothermic animals extensively produce rest lines that are annual or seasonal, but endothermic animals can also do so, even when temperature and food supply are maintained through the year^{23,25}. Moreover, different bones in the same skeleton can have sharply differing numbers of growth lines^{11,24}; the distance between successive growth lines does not always decrease regularly in some taxa¹²; and some taxa

closely related to those with copious growth lines may show none at all²⁶. Obviously the whole issue of bone ‘growth lines’ is linked more to organ-specific developmental dynamics, species-specific life-history strategies, and population-level interactions with their environments, than to overall patterns of thermometabolic Gestalt¹¹.

Dinosaurian versus reptilian growth rates

Comparing these two histological lines of evidence, our own investigations of extant and extinct archosaurs (including birds), as well as a survey of the published literature on bone histology, reveal a dichotomy between those archosaurs related to crocodiles and those related to birds and dinosaurs (Fig. 1). This distinction can be traced back to the division of the two lineages at least by the Middle Triassic, over 230 million years ago. In the hadrosaur *Maiasaura*, for example, the kinds of tissue typically deposited throughout the cortex of the long bones imply growth rates which suggest that maturity was reached at about seven years old and seven metres long¹¹. This estimate is consistent with counts of growth lines in the largest long bones, and the amount of tissue deposited between successive lines is commensurate with expected annual growth at rates reflected by the given bone tissues¹¹. (However, like extant mammals and birds, and other extinct dinosaurs, *Maiasaura* appears to have been growing too rapidly to lay down a growth line in its first year.) Similar growth rates are now estimated even for large sauropod dinosaurs^{9,13,14} and large pterosaurs²⁷. Such rates are in sharp contrast to those assessed for crocodiles, even gigantic ones like the Cretaceous *Deinosuchus*, which could exceed eight metres in length but took 50 years to do so¹⁵ (Fig. 2).

These results support the hypothesis that large dinosaurs grew more like large extant birds and mammals than like extant reptiles, even large ones^{2,4,11,18}. And, if growth rates provide any indication of underlying basal metabolic rates, they suggest that these dinosaurs were not like typical ectotherms^{2,3,6,10,11}. Finally, although most extant birds reach maturity too rapidly to lay down growth lines^{28,29}, some extinct birds, such as the two-metre-tall Eocene neognath *Diatryma* and the three-metre-tall moa, *Dinornis*, produced growth lines in their long bones (Fig. 3), suggesting not that they were ectothermic, but that the production of these lines was merely a function of endogenous rhythms^{17,24,30}.

In contrast to large dinosaurs and pterosaurs, small ones apparently grew more slowly. Their long bone cortices were less well vascularized, the vessels were primarily longitudinal, and the bones may show more closely spaced growth lines. These data would seem to fit a model describing dinosaurs as essentially reptilian, perhaps with somewhat higher basal metabolic rates, and large dinosaurs would grow quickly merely by virtue of inertial homoiothermy. Against this quite reasonable model, however, small mammals and birds also deposit bone of relatively low vascularity, with predominantly longitudinal vessels and lines of arrested growth (LAGs)^{28,29},

but are clearly endothermic^{2,3}. Moreover, the cortical bone of small dinosaurs and pterosaurs is primarily fibro-lamellar (not lamellar-zonal as in crocodiles and other ectotherms), reflecting more rapid growth, and the tissues are typically far better vascularized than in typical reptiles. These differences in patterns seem to begin before hatching¹⁰. Furthermore, extant taxa of large adult size typically grow at rates absolutely higher than related smaller taxa, regardless of their physiologies¹, so this pattern would be expected in dinosaurs. Even the most basal dinosaurs, such as the 2.5-metre-long *Herrerasaurus* and the 1.5-metre-long *Coelophysis*, grew at high rates, suggesting that smaller dinosaurs (for example, *Scutellosaurus* and *Orodromeus*) grew at secondarily slow rates (Fig. 1).

How ancient birds grew

These generalizations relating bone-tissue type and growth rates, when seen in phylogenetic perspective, can provide insight into how ancient birds grew and how their small size evolved. Compared with nearly all of their immediate relatives among the theropod dinosaurs, the dromaeosaurs and troodontids, the first birds were small³¹ (Fig. 4). Close bird relatives generally ranged from 1.1 to 3.0 metres in body length, with femur lengths of 14 to 30 centimetres. An exception is the recently discovered *Microraptor*³², which—if truly adult—is small enough (femur length 5.3 centimetres) to demonstrate that non-avian maniraptorans of suitable size for avian ancestry indeed existed. By comparison, *Archaeopteryx*, the most basal bird, reached a body length of approximately 50–60 centimetres (femur length about 3.2–7.0 centimetres), smaller than its feathered non-avian dinosaur relatives³³.

Previous analyses³⁴ of the bone tissue histology of two types of basal Mesozoic birds reported generally low vascularization and even a nearly avascular cortex in two enantiornithine femora, with several LAGs (Fig. 3). The lack of vascular canals and the persistence of LAGs are features commonly found in ectothermic extant reptiles and amphibians^{22,30}. Such features prompted the inference³⁴ that these basal birds had not fully attained avian physiological levels and may even have been intermediate between 'typical' reptilian metabolism and endothermy. Analyses of *Hesperornis* and *Ichthyornis*, two Cretaceous birds closer to extant forms, showed higher vascularity and an absence of LAGs, suggesting gradual attainment of the fully endothermic metabolism of extant birds³⁵. Other workers³⁶ suggested from similar histological studies that the basal bird *Confuciusornis* was endothermic, but that its endothermy had evolved independently of neornithine (extant) birds.

The inferences drawn from these histological observations are reasonable if one considers the distributions of such features only in extant vertebrates (ectothermic reptiles versus endothermic mammals and birds). However, the same evidence in a phylogenetic and ontogenetic context, including extinct forms (Fig. 4), suggests a different picture for basal birds^{5,17} (Fig. 3). Like all other theropod outgroups of birds studied, ranging from troodontids and tyrannosaurs out to allosaurs, coelophyids and even herrerasaurs, the small theropods from which birds evolved had well vascularized bone tissues¹⁶ (K.P. *et al.* unpublished observations). The long bones of the hindlimbs in particular consist almost entirely of tissues displaying a fibro-lamellar pattern⁴. This pattern, which also predominates in large extant birds and mammals, typically grows continuously at rates of 10–60 micrometres per day, much faster than the tissues of 'typical' extant reptiles^{8,22,37}. LAGs are observed throughout the cortices of the long bones of most non-avian theropods and other dinosaurs, but they are also known in the long bones of mammals and some extant birds, especially in the outermost cortex where growth slows^{28,29}. Thus, in these taxa they may represent little more than evidence of environmentally mediated endogenous patterns of cyclical growth with no general physiological implications²⁴. Slow-growing bone does not necessarily imply low metabolic rates, as can be seen from humans, whose bones grow much more slowly than those of most mammals, but who are obviously endothermic.

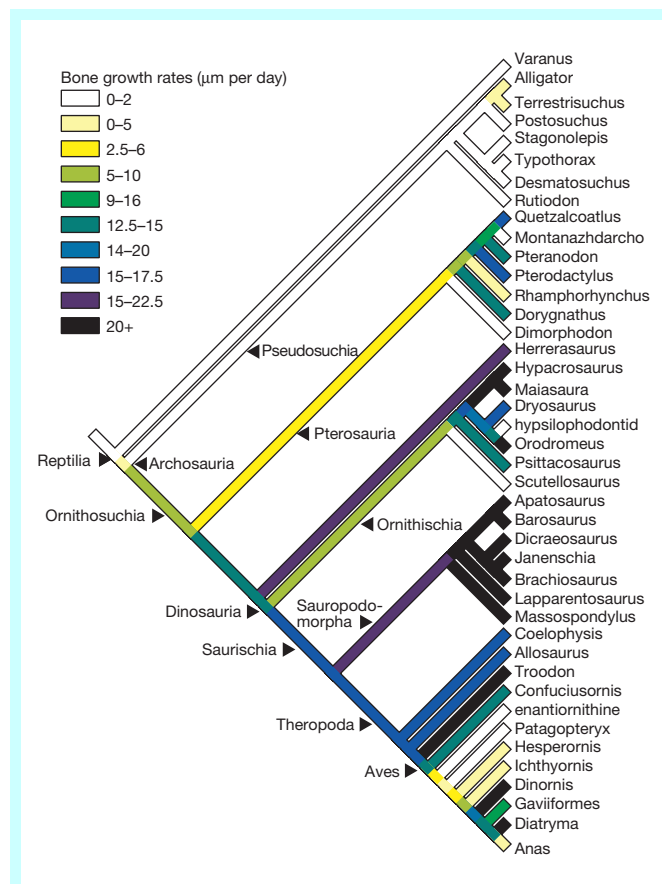


Figure 1 Cladogram of archosaurian taxa, after various sources, with subadult long-bone growth rates. Rates are estimates based on histological structure, vascularity, and vascular patterns of cortical tissues of subadult femora and tibiae, compared to the same types of tissues of known growth rate in the mallard, ostrich and emu^{19,37}. Growth rate was optimized as a continuous-state character (with values adjusted from the key) on a standard phylogeny of archosaurs using squared-change parsimony in MacClade version 3.08.

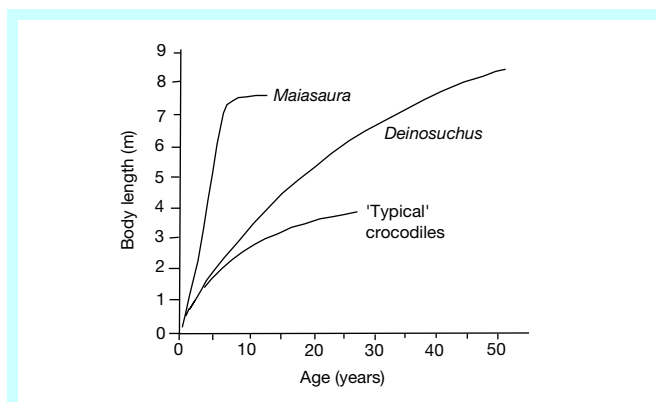


Figure 2 Comparative growth histories of the hadrosaurian dinosaur *Maiasaura*¹¹ and the giant Cretaceous crocodile *Deinosuchus*, with 'typical' crocodiles for comparison¹⁵. *Deinosuchus* grows slightly more rapidly and extends its active growth curve more than typical crocodiles, but the curve for *Maiasaura*, a typical large dinosaur with respect to its growth profile, is qualitatively different, reaching approximately the same adult size in about seven years, rather than 40 years or more. Although data on partial growth series have been collected for some dinosaurs^{9,10,12–14,16,18,20}, none but *Maiasaura* is currently represented from embryo to adult.

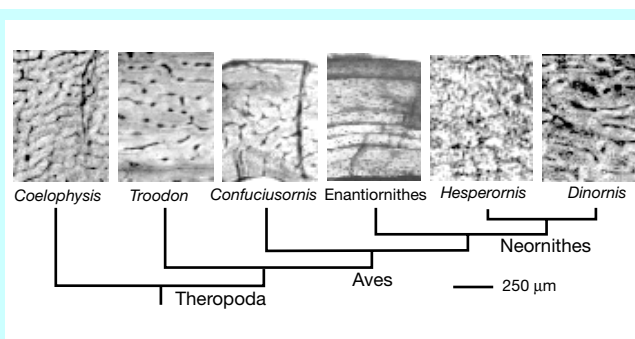


Figure 3 Transverse mid-shaft thin-sections of cortical bone of representative theropod taxa, showing the evolution of tissue from basal theropods to crown-group birds. The anomalous slow growth patterns in enantiornithines and in the external cortex of *Confuciusornis* are reflected by the nearly avascular, non-fibrolamellar tissue. (Enantiornithes from ref. 34; *Hesperornis* from ref. 35.)

If avian precursors had well vascularized, fast-growing bone, how can the relatively avascular, slow-growing patterns seen in some basal birds be explained? One possibility is that the limited tissue samples of enantiornithines may simply represent old individuals, in which growth has nearly ceased. Birds, like other theropods, have a thin cortex in which 75 per cent or more of earlier deposited tissue is resorbed, leaving no record in adults of presumably faster-growing juvenile tissues^{16,17}.

How birds got small

We propose an alternative hypothesis that takes into account both ontogeny and phylogeny. Preliminary histological observations of the basal bird *Confuciusornis*³⁶ (Fig. 3) do not suggest overall low growth rates, and a thorough study of this bird confirms active growth during a large part of its ontogeny at least (K.P. *et al.* unpublished observations). Conversely, bone tissue patterns indicate

that some basal birds, such as some enantiornithines and perhaps *Patagopteryx*, grew more slowly than extant birds of similar size^{34,35}. Thus they reached a much smaller adult size than other theropods, but they still took longer to grow to this smaller size than do extant birds. Their bone cortices suggest that they could have accomplished this by shortening the early rapid-growth phase and extending the more adult-like, slower-growth phase seen in some enantiornithine bones^{5,17,34}.

Further evidence for differences in growth curves between basal and neornithine birds comes from estimates of growth rates in the long bones of *Confuciusornis*^{17,31} (K.P. *et al.* unpublished observations). If the tibia is accorded a relatively slow growth rate of 4–6 micrometres per day, as its moderately vascularized cortex and mostly longitudinally oriented vascular canals indicate by comparison with the growth curves of extant birds³⁷, it would have reached adult size in just over half a year. This estimate is commensurate with those for the humerus (26 weeks), radius (21–31 weeks), and femur (35–48 weeks), if allowances are made for the range of growth curves³⁷ reflected in their respective tissues¹⁷. In contrast, most extant birds of the size of *Confuciusornis* reach adult size in four to eight weeks³⁸, but the larger theropod dinosaurs must have taken much longer, even at higher growth rates.

This hypothesis presumes that cortical bone tissue types that grow at specific rates in extant animals grew at similar rates in extinct forms. If these basal birds grew at such moderate overall rates (comprising a short phase of rapid growth and a protracted later phase of slower growth), it would explain why the known specimens of *Archaeopteryx* have a twofold size range, yet fit a single allometric growth trajectory³⁹.

A strict test for heterochrony⁴⁰, often invoked in connection with the origin of birds^{5,17}, requires complete comparative ontogenies of all forms under study, which is not possible in extinct tetrapods. However, the distribution of primitive and derived features in a phylogenetic series of related taxa provides a secondary test that reflects the direction of evolutionary changes and allows inferences about specific developmental mechanisms.

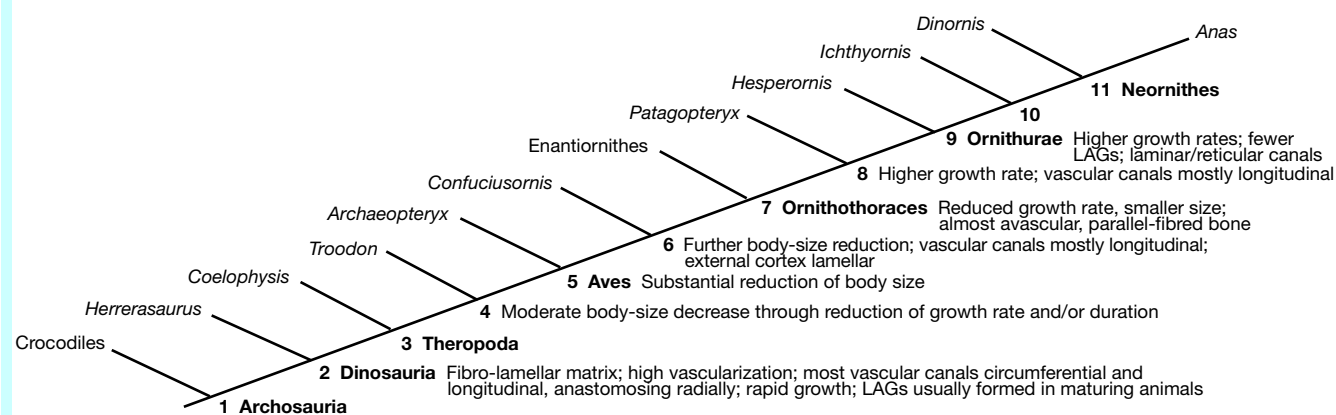


Figure 4 Cladogram of representative birds and other theropod dinosaurs, using crocodiles as an outgroup, to show the evolution of the growth strategies suggested by histological patterns, sampled in all taxa here except *Archaeopteryx*. These characters are treated as synapomorphies at the following nodes. **Node 1** (crocodiles): primarily parallel-fibred matrix; extensive LAGs; moderate vascularization; most vascular canals longitudinal; moderately slow growth. These features are common to Pseudosuchia basally but are not assumed here to be plesiomorphic for Archosauria. **Node 2** (Dinosauria): fibro-lamellar matrix extensively developed; dense vascularization with circular, longitudinal and radial canals; LAGs usually formed in maturing animals; rapid growth. **Nodes 4 and 5** (comprising the origin of birds and flight): as for Dinosauria; lamellar external cortex; reduction of body size through slowing of growth rate and/or shorter ontogeny, as reflected in bone tissues. **Node 6** (some features presumably apply to *Archaeopteryx*): longitudinal canals predominate with few anastomoses; body size reduced from theropod condition; lamellar external cortex and mostly longitudinal vascular canals in adults. **Node 7** (toward more advanced fliers): further reduction of vascularization; parallel-fibred or lamellar, almost non-vascular bone in adults; basally, further reduction of body size through further truncation of the early phase of fast growth. **Node 8**: fibro-lamellar cortex with longitudinal canals and few anastomoses; external cortex lamellar; fewer LAGs (?); higher growth rate reflected by faster-growing bone during a larger percentage of ontogeny. **Nodes 9–11** (toward the 'modern' condition): fibro-lamellar cortex with laminar or reticular vascularization in large birds; high to very high growth rates; fast-growing bone almost entirely resorbed in adults of small size; few or no LAGs.

If birds had simply reduced adult size by becoming paedomorphic (*sensu* ref. 40), they would have retained a juvenile shape at this reduced adult size (as, for example, if humans were to slow development so as to achieve adulthood and sexual maturity at a shape and size now characteristic of five-year-olds). Evidence for this pattern includes the persistence of features associated with juveniles of related taxa, and sometimes in the delay in, or absence of, fusion of skeletal elements commonly fused in the adults of outgroup taxa. However, in adult basal birds the forms and proportions of juvenile ancestors do not reappear. Rather, they show a derived development of proportional trends such as the further elongation of lower limbs, arms and hands, the beginnings of fusion of some skeletal elements (carpometacarpus, tibiotarsus, tarsometatarsus and pygostyle), and the further development of feathers^{5,17,31}. These patterns instead suggest a combination of overall dwarfing of adult size, accompanied by a form of peramorphosis of many skeletal forms and proportions. In other words, mosaic developmental changes carried some features of shape to more derived states even as body size was reduced⁴⁰. These changes are distinct from strict proportional dwarfism, because some shape parameters (such as the elongated arms and feathers) experienced positive allometry even as size decreased (acceleration *sensu* ref. 40). Reduction in body size was almost certainly associated with a reduction of the time needed to reach adult size. The relative elongation of forelimbs and feathers, coincident with a phyletic reduction in adult size, would have been advantageous to the inception of flight by decreasing wing loading and improving the power-to-weight ratio.

A boost in growth rates for later birds

Birds more closely related to extant groups (Neornithes, Fig. 4) gradually acquired the histological features of typical extant bird bone³⁵. Vascularity increased and LAGs were generally reduced or lost, although they persist at least in the outer cortex of some extant and fossil crown-group birds^{28,29}. The return to predominantly fast-growing bone tissue in more derived small birds (Neornithes and their relatives) suggests a trend to reach adult size quickly, within a shorter maturation time than in basal birds and non-avian theropods. (This pattern can be seen only in analysis of the 90 per cent of the growth trajectory comprising rapidly growing bone that is resorbed throughout ontogeny.) The acceleration of growth rates signalled further changes in life-history strategy that may have facilitated the rapid growth of some taxa to fledging size³⁸. Large ground birds such as ratites have converged in body size and bone histology with many Mesozoic non-avian theropods (Fig. 3). This seems to have resulted in large part from a secondary extension of the rapid-growth phase of their developmental trajectory that allows them to attain a large size quickly¹⁹.

Note added in proof: We are delighted that the independent results of Erickson *et al.*⁴¹ (this issue) so closely reflect and complement our own. Note that our measurements are linear, whereas theirs are extrapolated mass estimates, which accounts for many differences in the curves. Though more work is needed, we conclude that our independent studies generally characterize dinosaur growth strategies with substantial accuracy. □

- Case, T. J. Speculations on the growth rate and reproduction of some dinosaurs. *Paleobiology* **3**, 320–328 (1978).
- Reid, R. E. H. In *The Complete Dinosaur* (eds Farlow, J. O. & Brett-Surman, M. K.) 403–413 (Indiana Univ. Press, Bloomington, 1997).
- Reid, R. E. H. In *The Complete Dinosaur* (eds Farlow, J. O. & Brett-Surman, M. K.) 449–473 (Indiana Univ. Press, Bloomington, 1997).
- Ricqlès, A. de in *A Cold Look at the Warm-blooded Dinosaurs* (eds Thomas, R. D. K. & Olson, E. C.) 103–139 (Westview, Boulder, 1980).
- Ricqlès, A. de. L'origine dinosaurienne des oiseaux et de l'endothermie avienne: les arguments histologiques. *Année Biol.* **39**, 69–100 (2000).
- Spotila, J. R. In *A Cold Look at the Warm-blooded Dinosaurs* (eds Thomas, R. D. K. & Olson, E. C.) 233–252 (Westview, Boulder, 1980).
- Dunham, A. E., Overall, K. L., Porter, W. P. & Forster, C. A. Implications of ecological energetics and biophysical and developmental constraints for life-history variation in dinosaurs. *Geol. Soc. Am. Sp. Pap.* **238**, 1–19 (1989).

- Ricqlès, A. de, Meunier, F. J., Castanet, J. & Francillon-Vieillot, H. in *Bone, Vol. 3: Bone Matrix and Bone Specific Products* (ed. Hall, B. K.) 1–78 (CRC, Boca Raton, 1991).
- Rimblot-Baly, F., Ricqlès, A. de & Zylberberg, L. Analyse paléohistologique d'une série de croissance partielle chez *Lapparentosaurus madagascariensis* (Jurassique moyen). *Ann. Paléontol.* **81**, 49–86 (1995).
- Horner, J. R., Padian, K. & Ricqlès, A. J. de Comparative osteohistology of some embryonic and perinatal archosaurs: phylogenetic and behavioral implications for dinosaurs. *Paleobiology* **27**, 39–58 (2001).
- Horner, J. R., Ricqlès, A. J. de & Padian, K. The bone histology of the hadrosaurid dinosaur *Maiasaura peeblesorum*: growth dynamics and physiology based on an ontogenetic series of skeletal elements. *J. Vert. Paleontol.* **20**, 109–123 (2000).
- Chinsamy, A. Bone histology and growth trajectory of the prosauropod dinosaur *Massospondylus carinatus* Owen. *Mod. Geol.* **18**, 319–219 (1993).
- Curry, K. A. Ontogenetic histology of *Apatosaurus* (Dinosauria: Sauropoda): New insights on growth rates and longevity. *J. Vert. Paleontol.* **19**, 654–665 (1999).
- Sander, P. M. Longbone histology of the Tendaguru sauropods: implications for growth and biology. *Paleobiology* **26**, 466–488 (2000).
- Erickson, G. M. & Brochu, C. A. How the 'terror crocodile' grew so big. *Nature* **398**, 205–206 (1999).
- Varricchio, D. J. Bone microstructure of the Upper Cretaceous theropod dinosaur *Troodon formosus*. *J. Vert. Paleontol.* **13**, 99–104 (1993).
- Ricqlès, A. de, Padian, K. & Horner, J. R. in *Perspectives on the Origin and Early Evolution of Birds* (ed. Gauthier, J. A.) (Yale Univ. Press, New Haven, 2001) (in the press).
- Varricchio, D. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 282–288 (Academic, San Diego, 1997).
- Castanet, J., Rogers, K. C., Cubo, J. & Boisard, J.-J. Periosteal bone growth rates in extant ratites (ostrich and emu). Implications for assessing growth in dinosaurs. *C.R. Acad. Sci. III* **323**, 543–550 (2000).
- Erickson, G. M. & Tumanova, T. A. Growth curve of *Psittacosaurus mongoliensis* Osborn (Ceratopsia: Psittacosauridae) inferred from long bone histology. *Zool. J. Linn. Soc.* **130**, 551–566 (2000).
- Amprino, R. La structure du tissu osseux envisagée comme expression de différences dans la vitesse de l'accroissement. *Arch. Biol.* **58**, 315–330 (1947).
- Castanet, J., Francillon-Vieillot, H., Meunier, F. J. & de Ricqlès, A. in *Bone, Vol. 7: Bone Growth B* (ed. Hall, B. K.) 245–283 (CRC, Boca Raton, 1993).
- Padian, K. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 288–291 (Academic, San Diego, 1997).
- Horner, J. R., Ricqlès, A. J. de & Padian, K. Variation in skeletochronological indicators of the hadrosaurid dinosaur *Hypacrosaurus*: implications for age assessment of dinosaurs. *Paleobiology* **25**, 295–304 (1999).
- Klevezal, G. A. *Recording Structures of Mammals: Determination of Age and Reconstruction of Life History* (Balkema, Rotterdam, 1996).
- Chinsamy, A., Rich, T. & Vickers-Rich, P. Polar dinosaur bone histology. *J. Vert. Paleontol.* **18**, 385–390 (1998).
- Ricqlès, A. de, Padian, K., Horner, J. R. & Francillon-Vieillot, H. Paleohistology of the bones of pterosaurs (Reptilia: Archosauria): anatomy, ontogeny, and biomechanical implications. *Zool. J. Linn. Soc.* **129**, 349–385 (2000).
- Van Soest, R. W. M. & Van Utrecht, W. L. The layered structure of bones of birds as a possible indication of age. *Bijdr. tot de Dierkde.* **41**, 61–66 (1971).
- Klevezal, G. A., Kallar Sallas, A. V. & Kirpichev, S. P. Determination of age in birds by layers in the periosteal bone. *Zool. Zh.* **51**, 1726–1730 (1972).
- de Buffrénil, V. & Castanet, J. Age estimation by skeletochronology in the Nile Monitor (*Varanus niloticus*), a highly exploited species. *J. Herpetol.* **34**, 414–424 (2000).
- Chiappe, L. M., Ji S.-A., Ji Q., & Norell, M. A. Anatomy and systematics of the Confuciusornithidae (Theropoda: Aves) from the Late Mesozoic of northeastern China. *Bull. Am. Mus. Nat. Hist.* **242**, 1–89 (1999).
- Xu, X., Zhou, Z. & Wang, X. The smallest known non-avian theropod dinosaur. *Nature* **408**, 705–708 (2001).
- Ji Q., Currie, P. J., Norell, M. A. & Ji, S.-A. Two feathered dinosaurs from northeastern China. *Nature* **393**, 753–761 (1998).
- Chinsamy, A., Chiappe, L. M. & Dodson, P. Mesozoic avian bone microstructure: physiological implications. *Paleobiology* **21**, 561–574 (1995).
- Chinsamy, A., Martin, L. D. & Dodson, P. Bone microstructure of the diving *Hesperornis* and the volant *Ichthyornis* from the Niobrara Chalk of western Kansas. *Cret. Res.* **19**, 225–235 (1998).
- Zhang, F., Hou, L. & Ouyang, L. Osteological microstructure of *Confuciusornis*: preliminary report. *Vert. Palasiatica* **36**, 126–135 (1998).
- Castanet, J., Grandin, A., Abourachid, A. & Ricqlès, A. de. Expression de la dynamique de croissance de l'os périostique chez *Anas platyrhynchos*. *C.R. Acad. Sci.* **319**, 301–308 (1996).
- Starck, J. M. & Ricklefs, R. E. (eds) *Avian Growth and Development: Evolution within the Altricial-Precocial Spectrum* (Oxford Univ. Press, New York, 1998).
- Houck, M. A., Gauthier, J. A. & Strauss, R. E. Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica*. *Science* **247**, 195–198 (1990).
- Alberch, P., Gould, S. J., Oster, G. F. & Wake, D. B. Size and shape in ontogeny and phylogeny. *Paleobiology* **5**, 296–317 (1979).
- Erickson, G. M., Curry Rogers, K. & Yerby, S. A. Dinosaurian growth patterns and rapid avian growth rates. *Nature* **412**, 429–433 (2001).

Acknowledgements

We thank M. Zelditch, J. Hutchinson and J. Cubo for comments, and K. Angielczyk for phylogenetic help. E. Lamm prepared the thin-sections. This work was supported by The Charlotte and Walter Kohler Charitable Trust, the CNRS, the Collège de France, the Miller Institute for Basic Research, and the Committee on Research of the University of California, Berkeley. This is University of California Museum of Paleontology Contribution No. 1741.

Correspondence should be addressed to K.P. (e-mail: kpadian@socrates.berkeley.edu).

The stability against freezing of an internal liquid-water ocean in Callisto

Javier Ruiz

Departamento de Geodinámica, Facultad de Ciencias Geológicas, and Seminar on Planetary Sciences, Universidad Complutense de Madrid, 28040 Madrid, Spain

The discovery of the induced magnetic field of Callisto—one of Jupiter's moons—has been interpreted^{1,2} as evidence for a subsurface ocean, even though the presence of such an ocean is difficult to understand in the context of existing theoretical models^{3–5}. Tidal heating should not be significant for Callisto, and, in the absence of such heating, it is difficult to see how this internal ocean could have survived until today without freezing^{1,6}. Previous work^{3,4} indicated that an outer ice layer on the ocean would be unstable against solid-state convection, which once begun would lead to total freezing of liquid water in about 10^8 years. Here I show that when a methodology⁷ for more physically reasonable water ice viscosities (that is, stress-dependent non-newtonian viscosities, rather than the stress-independent newtonian viscosities considered previously) is adopted, the outer ice shell becomes stable against convection. This implies that a subsurface ocean could have survived up to the present, without the need for invoking antifreeze substances or other special conditions.

The moment of inertia of Callisto, as determined from measurements taken by the Galileo spacecraft, indicates that the interior of the satellite is probably partly differentiated⁸. This interpretation is in accordance with the existence of an outer water layer (frozen, liquid, or both) less than 350 km thick, overlying an interior consisting essentially (if not totally) of an ice and rock mixture, but it does not constitute as complete a differentiation as that of Ganymede⁸. It is also consistent with a depth of less than 300 km from the surface to the ocean top, as estimated from the amplitude of the induced magnetic field².

The first detailed thermal models of icy satellites (see, for example, refs 9 and 10) did not take into account the possibility of solid-state convection. They assumed heat transfer through the outermost layers by thermal conduction only (which would permit ice melting, the complete differentiation of the bodies, and the existence of thick layers of liquid water inside the satellites). Later workers^{3–5} concluded that the outer ice layers of the larger icy satellites, such as Ganymede and Callisto, are unstable against convection. The viscosity of water ice was taken as newtonian—it was described as an exponential function of temperature, but without taking into account stresses or strain rates.

An internal ocean could avoid total freezing if the ice above it was more rigid than is usually assumed, so that convective heat flow was reduced^{4,6}. Although the possibility has been mentioned that the real, non-newtonian, behaviour of water ice could have the same effect⁴, this point has not been modelled. Another way to avoid the freezing of an internal ocean is the presence of substances that lower the melting point of ice^{1,4,11}. The inclusion of ammonia could decrease the melting point to ~ 176 K (see, for example, ref. 12), and the presence of chloride salts or sulphuric acid could lower it to about 210 K (ref. 13). We note that the pioneering work of ref. 9 considered the possible role of ammonia in relation to the existence of liquid water and differentiation in Callisto. Moreover, a possible decrease of the melting point of ice owing to the presence of ammonia would also increase the viscosity of the ice shell, which in turn would reduce the convective heat flow¹⁴.

In ductile deformation, the strain rate is proportional to the applied stress: $\dot{\epsilon} \propto \sigma^n$, where n depends on the creep mechanism, and therefore on stress, temperature and grain size. Viscosity is

newtonian (independent of $\dot{\epsilon}$ and σ) if $n = 1$ (diffusion creep). On the basis of modern experimental results and theoretical considerations, it has been proposed^{15–18} that, in the low-stress (~ 0.1 MPa) conditions inside icy satellites, grain boundary sliding ($n = 1.8$ in water ice I, which is the water ice polymorph that constitutes the outer layer of Callisto^{3,4}) could be the dominant creep mechanism, at least for grain sizes less than ~ 1 mm, instead of diffusion creep as was previously thought⁵. But low stress and warm temperatures favour the growth of ice crystals, and higher grain sizes increase the relative importance of dislocation creep ($n \approx 4$ in water ice I). For that reason, it could be necessary to consider both creep mechanisms when modelling of the interior of large icy satellites¹⁹. If grain boundary sliding and dislocation creep coexist, n takes effective values between 1.8 and 4. On the other hand, diffusion creep does not seem to be a relevant deformation mechanism under planetary conditions²⁰.

In a more realistic analysis of the stability against convection of a shell of water ice I above an internal ocean in Callisto—such as I report here—non-newtonian ($n > 1$) rheological behaviour needs to be taken into account. Any substances potentially present in the water ice are not considered in my calculations, as it is not known whether they occur in amounts sufficient to produce high-degree partial melting or to alter significantly the properties of the ice. The use of water ice implies an endmember model, in which the conditions for maintaining an ocean stable against freezing are more restrictive.

The stability against convection of a layer (in this case, Callisto's outer ice shell) can be estimated by means of the Rayleigh number defined at the layer base, Ra_{base} ; for non-newtonian viscosity, this is

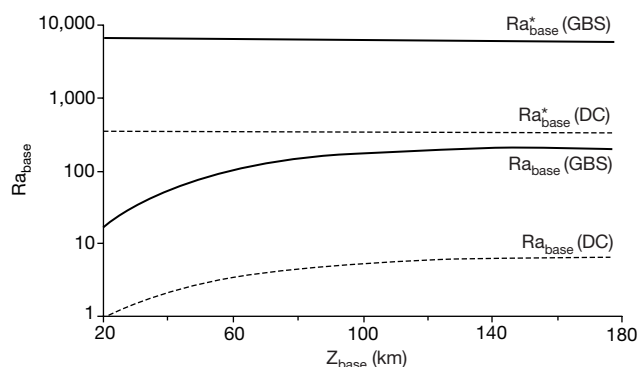


Figure 1 Stability against convection of Callisto's outer ice shell. Ra_{base} (the Rayleigh number at the base of the outer ice shell) and Ra_{base}^* (the critical value of Ra_{base}) are shown in terms of Z_{base} (the total thickness of the outer ice shell), for both grain boundary sliding (GBS) and dislocation creep (DC). Ra_{base}^* is nearly constant, and has been estimated from θ after ref. 7. For ice I, several terms involved in the calculation of Ra_{base} are functions of temperature: $\alpha = 1.56(T/250) \times 10^{-4} \text{ K}^{-1}$, and $\kappa = 1.47(250/T)^2 \times 10^{-6} \text{ m}^2 \text{ s}^{-1}$ (ref. 24), where $T = T_i$ for the convective state analysis¹⁷. The density of water ice I varies slightly with temperature and pressure, but the variations obtained from equation of state measurements for this substance are slight²⁸, therefore the adoption of a constant value does not alter the results significantly. With regard to experimentally established creep parameters: for grain boundary sliding^{15,18} $A = 9.80 \times 10^{-10} \text{ Pa}^{-n} \text{ mm}^p \text{ s}^{-1}$, $n = 1.8$, $p = 1.4$, and $Q = 49 \text{ kJ mol}^{-1}$ (Ra_{base} has been calculated for 0.1-mm grain size); and for dislocation creep²⁹ $A = 1.26 \times 10^{-19} \text{ Pa}^{-n} \text{ s}^{-1}$, $n = 4$, $p = 0$ and $Q = 61 \text{ kJ mol}^{-1}$. The creep of water ice I shows high-temperature regimes where enhanced creep rates occur. These regimes are associated with premelting effects¹⁸, or with recrystallization under stress³⁰, and have still higher values¹⁸ of Q ; they dominate at temperatures above ~ 240 – 260 K (different reports give different temperatures in this range^{18,29}). Higher values of Q reduce Ra_{base} , because—according to equations (1) and (6)— $Ra_{\text{base}} \propto A^{1/n} \exp[Q(2 - \Delta T/T_i)/nRT_i]$, and they simultaneously increase θ and Ra_{base}^* (ref. 7). Therefore, the use of the numerical values for creep parameters listed above places upper limits on Ra_{base} and lower limits on Ra_{base}^* , for both creep mechanisms—grain boundary sliding and dislocation creep.

given by⁷

$$Ra_{\text{base}} = \frac{\alpha g \rho h^{(n+2)/n} \Delta T}{\kappa^{1/n} b^{1/n}} \exp(\theta/n) \quad (1)$$

where α is the volumetric thermal expansion coefficient, g is the acceleration due to gravity (taken in general as the surface value, 1.24 m s^{-2} for Callisto), ρ is the density (930 kg m^{-3} for water ice I), h is the effective layer thickness, $\Delta T (= T_{\text{base}} - T_s)$ is the temperature difference between the base and the surface of the layer, κ is the thermal diffusion coefficient, b is a parameter that depends on creep mechanism and temperature, and n is a constant that depends on creep mechanism. $\theta (= Q\Delta T/RT_i^2)$ is the Frank–Kamenetskii parameter (which is related to the viscosity contrast through the layer caused by temperature differences), where Q is the activation enthalpy of creep deformation (which, because of experimental uncertainty, can be taken as the activation energy), R is the gas constant, and T_i is the adiabatic temperature (approximately constant) in the case of convection. If, for a definite value of θ , Ra_{base} does not reach a certain critical value Ra_{base}^* (estimated after ref. 7), a layer in thermal conductive equilibrium is stable against convection.

To solve equation (1), it is necessary to make a reasonable estimate of T_{base} , T_i and h . Before the onset of convection, heat is transmitted into Callisto's outer shell only by conduction. As there is no tidal heating in Callisto^{1,3,21}, I considered an outer shell in conductive thermal equilibrium, and heated from below by radioactive disintegration in the interior rocky fraction underneath the ocean. Under these conditions, the variation of temperature with depth can be described through Fourier's law,

$$dT = (F_z/k) dz \quad (2)$$

where F_z is the vertical heat flow at a depth z from surface, and k is the thermal conductivity. For ice I, k is a function of temperature ($k = k_0/T$, where k_0 is a constant with a value of 567 W m^{-1} (ref. 22)); also, in a spherical layer in energetic equilibrium heated from below, $F_z = r^2 F/(r - z)^2$, where r is the body radius, and F is the heat flow at the surface. If both expressions for k and F_z are substituted in equation (2), and the equation is then integrated from the surface to z_{base} (z_{base} is the total thickness of the outer ice shell), an equation for temperature at the base of the outer ice shell is obtained:

$$T_{\text{base}} = T_s \exp \left[\frac{r F z_{\text{base}}}{k_0 (r - z_{\text{base}})} \right] \quad (3)$$

where T_s is taken as 130 K (ref. 23). At the same time, if the Clapeyron slope for melting ice is taken as $-0.1063 \text{ K MPa}^{-1}$ (ref. 24), T_{base} (in K) can be approximated as

$$T_{\text{base}} = 273.16 - 0.1063 P_{\text{base}} \quad (4)$$

where P_{base} (in MPa) is the pressure at the outer shell base, given by integration from the surface to z_{base} of $\rho G M_z/(r - z)^2 dz$, where in turn G is the gravitational constant, and M_z is the mass of a sphere with radius $(r - z)$ (the total mass of Callisto is $2.5 \times 10^{23} \text{ kg}$). If equations (3) and (4) are resolved simultaneously, z_{base} and T_{base} are obtained. Because thermal conductivity is a function of temperature, h is not the real layer thickness z_{base} before the onset of convection, but an effective thickness¹⁷, which is defined for a situation of conductive thermal equilibrium with the constant thermal conductivity corresponding to T_i ; the convective analysis followed here assumes a cartesian geometry⁷, and for that reason $h = k_0 \Delta T/FT_i$.

Convection on icy satellites would involve a stagnant-lid regime²⁶, in which the upper part of an outer shell does not participate in the convective movement. The temperature contrast across the lower boundary layer of the convecting region, $T_{\text{base}} - T_i$, can be taken as the rheological temperature scale²⁷, $\sim \Delta T/\theta$, and for that reason T_i can be estimated approximately from:

$$T_{\text{base}} - T_i \approx RT_i^2/Q \quad (5)$$

The effective viscosity for a non-newtonian rheology can be described by $\eta = (d^p/A)\sigma_{\text{II}}^{1-n}\exp(Q/RT)$, where d is the grain size, p and A are experimentally established constants, and σ_{II} is the second invariant of the deviatoric stress tensor. In some work on solid-state convection^{7,27}, the use of the Frank–Kamenetskii approximation ($\eta = b\sigma_{\text{II}}^{1-n}\exp(-QT/RT_i^2)$) as the value of the effective viscosity has been proposed. If the condition is imposed that both expressions give the same η value in $T = T_i$, b can be estimated in T_i from:

$$b = (d^p/A) \exp(2Q/RT_i) \quad (6)$$

I performed the analysis for both grain boundary sliding and dislocation creep, according to the discussion above. Grain boundary sliding is a grain-size-sensitive creep mechanism, in which smaller grain sizes diminish the b value, thus conversely increasing Ra_{base} . For that reason, I have taken a grain size of 0.1 mm: the existence of smaller grain sizes does not seem a realistic possibility for the interiors of large icy satellites¹⁷, and therefore this value implies a reasonable upper limit in the estimation of Ra_{base} for grain boundary sliding. On the other hand, dislocation creep is independent of grain size, and for that reason $p = 0$.

The results are given in Fig. 1; they are shown as Ra_{base} in terms of z_{base} and compared with the corresponding Ra_{base}^* values, for both grain boundary sliding and dislocation creep. z_{base} increases gradually with time as F diminishes because of power loss of the internal heat sources. If the present surface heat flow is taken as $\sim 3.6 \text{ mW m}^{-2}$, then today $z_{\text{base}} \approx 105 \text{ km}$. The value of $\sim 3.6 \text{ mW m}^{-2}$ is the mean value of the heat flow interval estimated as a function of the degree of differentiation and of the rock models for the rocky fraction of Callisto²³: it varies from 3.3 mW m^{-2} (Callisto undifferentiated) to 3.9 mW m^{-2} (Callisto completely differentiated). On the other hand, $F \approx 2 \text{ mW m}^{-2}$ and $z_{\text{base}} = 176.5 \text{ km}$ corresponds to the minimal melting point as a function of the pressure of ice (at 251.1 K and 207 MPa, ref. 24).

The data in Fig. 1 show that Callisto's conductive ice shell is at present stable against solid-state convection—because the values of Ra_{base} are always much lower than those of Ra_{base}^* , for both grain boundary sliding and dislocation creep. These results are important, because they imply that a conductive outer shell would also have been stable against convection in the past.

This calculated stability of an internal ocean and a non-convecting ice shell could be consistent with the observed geology of Callisto. It could be argued that the heavily cratered surface of Callisto, and the lack of clear geological modifications originating from under the surface, are evidence against the existence of a subsurface ocean¹; but if the icy shell is stable and convectively inactive, and if the cold, brittle lithosphere is very thick, geological inactivity of the shell (compared to those of the tidally heated^{4,21} jovian satellites Europa and possibly Ganymede) can be understood.

Finally, according to these results for Callisto, a consequence of the non-newtonian nature of water ice I could be that it is more difficult than previously thought to obtain convection in the external layers of icy satellites. If this is correct, the existence of oceans under ice in great icy satellites could be a common phenomenon. $\square$

Received 13 November 2000; accepted 17 May 2001.

1. Khurana, K. K. *et al.* Induced magnetic fields as evidence for subsurface oceans in Europa and Callisto. *Nature* **395**, 777–780 (1998).
2. Zimmer, C., Khurana, K. K. & Kiverson, M. G. Subsurface oceans on Europa and Callisto: Constraints from Galileo magnetometer observations. *Icarus* **147**, 329–347 (2000).
3. Reynolds, R. T. & Cassen, P. M. On the internal structure of the major satellites of the outer planets. *Geophys. Res. Lett.* **6**, 121–124 (1979).
4. Cassen, P. M., Peale, S. J. & Reynolds, R. T. in *Satellites of Jupiter* (ed. Morrison, D.) 93–128 (Univ. Arizona Press, Tucson, 1982).
5. Schubert, G., Spohn, T. & Reynolds, R. T. in *Satellites* (eds Burns, J. A. & Matthews, M. S.) 224–292 (Univ. Arizona Press, Tucson, 1986).
6. Showman, A. P. & Malhotra, R. The Galilean satellites. *Science* **286**, 77–84 (1999).
7. Solomatov, V. S. Scaling of temperature- and stress-dependent viscosity convection. *Phys. Fluids* **7**, 266–274 (1995).
8. Anderson, J. D. *et al.* Distribution of rock, metals, and ices in Callisto. *Science* **280**, 1573–1576 (1998).

9. Lewis, J. S. Satellites of the outer planets: thermal models. *Science* **172**, 1127–1128 (1971).
10. Consolmagno, G. J. & Lewis, J. S. The evolution of icy satellite interiors and surfaces. *Icarus* **34**, 280–293 (1978).
11. Stevenson, D. J. An ocean within Callisto? *Eos* **79**, F535 (1998).
12. Kargel, J. S. Ammonia-water volcanism on icy satellites: Phase relations at 1 atmosphere. *Icarus* **100**, 556–574 (1992).
13. Kargel, J. S. *et al.* Europa's crust and ocean: origin, composition, and the prospects for life. *Icarus* **148**, 226–265 (2000).
14. Sotin, C., Grasset, O. & Beauchesne, S. in *Solar System Ices* (eds Schmitt, B., De Bergh, C. & Festou, M.) 79–96 (Kluwer Academic, Dordrecht, 1998).
15. Goldsby, D. L. & Kohlstedt, D. L. Grain boundary sliding in fine-grained ice I. *Scripta Mater.* **37**, 1399–1406 (1997).
16. Pappalardo, R. T. *et al.* Geological evidence for solid-state convection in Europa's ice shell. *Nature* **391**, 365–368 (1998).
17. McKinnon, W. B. Convective instability in Europa's floating ice shell. *Geophys. Res. Lett.* **26**, 951–954 (1999).
18. Goldsby, D. L. & Kohlstedt, D. L. Superplastic deformation of ice: Experimental observations. *J. Geophys. Res.* (in the press).
19. Durham, W. B., Kirby, S. H. & Stern, L. A. The rheology of ice I at low stress and elevated confining pressure. *J. Geophys. Res.* (in the press).
20. Goldsby, D. L. & Kohlstedt, D. L. Flow of ice I by dislocation, grain boundary sliding, and diffusion processes. *Lunar Planet. Sci. XXVIII*, 429–430 (1997).
21. Showman, A. P. & Malhotra, R. Tidal evolution into Laplace resonance and the resurfacing of Ganymede. *Icarus* **127**, 93–111 (1997).
22. Klinger, J. Influence of a phase transition of the ice on the heat and mass balance of comets. *Science* **209**, 271–272 (1980).
23. Mueller, S. & McKinnon, W. B. Three-layered models of Ganymede and Callisto: Compositions, structures, and aspects of evolution. *Icarus* **76**, 437–464 (1988).
24. Kirk, R. L. & Stevenson, D. J. Thermal evolution of a differentiated Ganymede and implications for surface features. *Icarus* **69**, 91–134 (1987).
25. Davies, M. E. *et al.* The control networks of the Galilean satellites and implications for global shape. *Icarus* **135**, 372–376 (1998).
26. McKinnon, W. B. in *Solar System Ices* (eds Schmitt, B., De Bergh, C. & Festou, M.) 525–550 (Kluwer Academic, Dordrecht, 1998).
27. Solomatov, V. S. & Moresi, L.-N. Stagnant lid convection on Venus. *J. Geophys. Res.* **101**, 4737–4753 (1996).
28. Lupo, M. J. & Lewis, J. S. Mass-radius relationships in icy satellites. *Icarus* **40**, 157–170 (1978).
29. Durham, W. B., Kirby, S. H. & Stern, L. A. Creep of water ices at planetary conditions: A compilation. *J. Geophys. Res.* **102**, 16293–16302 (1997).
30. Kamb, B. in *Flow and Fracture of Rocks* (eds Heard, H. C., Borg, I. Y., Carter, N. L. & Raleigh, C. B.) 211–241 (AGU Geophys. Monogr. Ser. 16, American Geophysical Union, Washington DC, 1972).

Acknowledgements

I thank F. Anguita, O. Prieto, J. Ruiz, R. Tejero and A. Torices for suggestions, K. Bennett and J. Kargel for comments on the manuscript, and W. Durham and D. Goldsby for preprints of work in press.

Correspondence and requests for materials should be addressed to the author (e-mail: jaruiz@eucmax.sim.ucm.es).

Evidence for recent climate change on Mars from the identification of youthful near-surface ground ice

John F. Mustard, Christopher D. Cooper & Moses K. Riffkin

Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA

Ground ice in the crust and soil may be one of the largest reservoirs of water on Mars^{1–3}. Near-surface ground ice is predicted to be stable at latitudes higher than 40° (ref. 4), where a number of geomorphologic features indicative of viscous creep and hence ground ice have been observed⁵. Mid-latitude soils have also been implicated as a water-ice reservoir⁶, the capacity of which is predicted to vary on a 100,000-year timescale owing to orbitally driven variations in climate⁷. It is uncertain, however, whether near-surface ground ice currently exists at these latitudes, and how it is changing with time. Here we report observational evidence for a mid-latitude reservoir of near-surface water ice occupying the pore space of soils. The thickness of the ice-occupied soil reservoir

(1–10 m) and its distribution in the 30° to 60° latitude bands indicate a reservoir of $(1.5–6.0) \times 10^4 \text{ km}^3$, equivalent to a global layer of water 10–40 cm thick. We infer that the reservoir was created during the last phase of high orbital obliquity less than 100,000 years ago, and is now being diminished.

Using high-resolution images from the Mars Orbiter Camera (MOC) on the Mars Global Surveyor (MGS) spacecraft^{8,9}, we have identified and mapped on a global scale a unique, young terrain on Mars that exhibits a morphology consistent with a material that has been cemented and then partially dissected (cut into hills and valleys) or disaggregated. The terrain (Fig. 1) is recognized on the basis of the following criteria: (1) a smooth, intact surface is present; (2) the smooth material is broken up or dissected somewhere in the

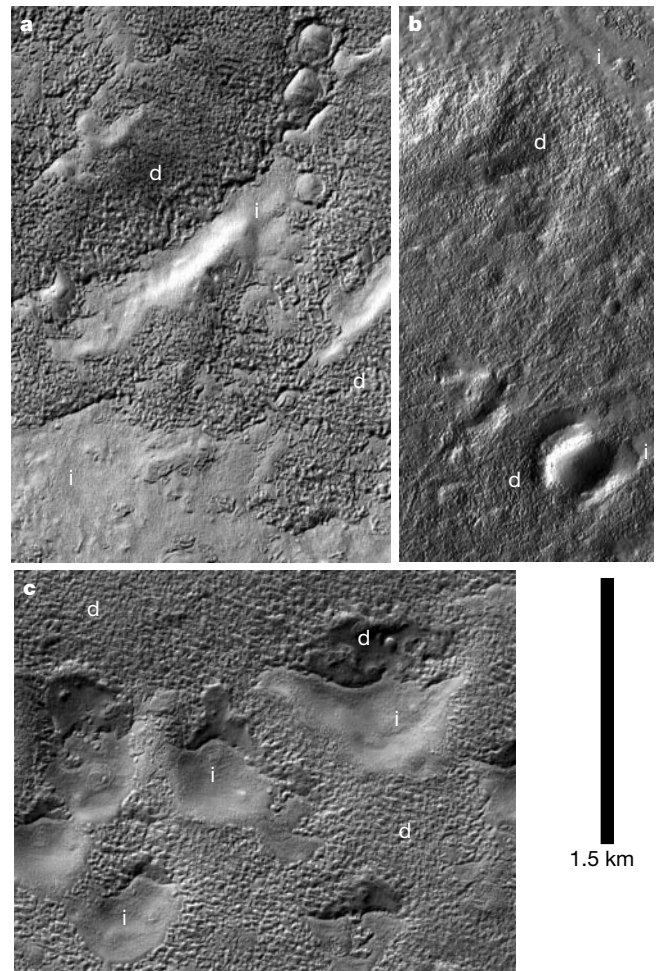


Figure 1 Examples of terrain interpreted to contain near-surface ground ice. The scale is the same for all images and in each image 'i' indicates intact terrain and 'd' indicates dissected (eroded) terrain. Illumination is from below and the nearest pole is towards the top of each image. **a**, A large expanse of intact terrain with incipient dissection into numerous steep-walled pits and troughs a few metres in size. The layer thickness is estimated to be 10 m. **b**, The amount of dissection is greater and clearly demonstrates the lack of preferred orientation and pattern in the dissected terrain. Here the layer thickness is estimated to be 2 m. **c**, The dissected layer in this region, estimated to be 5 m thick, is completely removed in places, revealing the substrate beneath. In these three examples, the density of pits ranges from sparse to abundant. The pitted surface grades into a knobby terrain which further grades into a rough surface. In areas of complete degradation a relict surface is revealed. The gradations between the smooth, pitted, knobby and rough surfaces imply a process where the smooth unit disaggregates and is removed to reveal a pre-existing surface. **a**, Image M0400837 (43.40°N, 240.21°W), $i = 55.31^\circ$; **b**, image SP253404 (34.35°N, 356.00°W), $i = 74.67^\circ$; **c**, image FHA01450 (43.71°S, 239.61°W), $i = 75.49^\circ$.

scene; (3) the pitted or hummocky terrain resulting from dissection is distinct from dunes, yardangs (ridges formed by wind erosion), or other aeolian (wind-driven) features.

We have examined over 8,000 of the narrow-angle MOC images that were acquired prior to August 1999 with a spatial resolution of less than 10 m per pixel. The images were classified by the presence or absence of the terrain as described above and the results are shown in Fig. 2. The eroded terrain is concentrated in two latitude bands: 30° to 70° N and 25° to 65° S. Few occurrences are found between $\pm 30^\circ$ latitude and at latitudes poleward of $\pm 60^\circ$. The total number of images examined and the fraction of images showing dissection, as a function of latitude, are shown in Fig. 3. Two distinct and remarkably similar distributions are observed: they are slightly asymmetric with a steep rise in number from equatorial latitudes but with extended occurrences towards polar latitudes.

Detailed analysis of many of the individual images with this terrain reveals the following general characteristics (Fig. 1). Where this terrain is intact, it forms a mantle of uniform thickness that is superimposed or draped across the topography muting the morphology. This is distinct from mantling in equatorial latitudes where such mantles fill in topographic lows leaving the rough textured surfaces on ridges and high points. The intact unit is very uniform in albedo, appears smooth at the resolution of the MOC images (typically 3 m per pixel), and fresh impact craters of any size are very rare. Pits and remnant ridges exhibit no apparent preferred orientations and the total absence of any layering indicates a lack of internal structure to the mantle layer. By using the shadows adjacent to pits along the azimuth direction of the sun we estimate the thickness of the deposit to range between 1 m and 10 m. Within a given image, estimates of layer thickness are consistent, indicating a discrete thickness to the layer in any given region. There is a conspicuous absence of mobile materials such as dunes associated with this unit; dunes are commonly observed elsewhere on Mars at MOC resolution⁹. Where the layer is completely removed (Fig. 1c), a more stable or solid surface remains apparently unaffected by the process of dissection. The morphology indicates the presence of a mantling material that has sufficient strength to support short, steep slopes.

The material comprising the majority of the mantle is probably dust. There are two models for the existence of a competent surface

mantle dominated by dust: a chemical cement or cementation with ground ice. A weakly cemented regolith was observed at the Viking landing sites, in which sulphate salts were considered to be the most likely cement¹⁰. Surface-atmosphere exchange of adsorbed water provides a mechanism to mobilize highly soluble magnesium sulphate on geological timescales, leading to a cemented mantle. However, an increase in the thermal inertia relative to terrains dominated by loose dust¹⁰, or spectroscopic properties that indicate sulphate-cemented soils¹¹ associated with the distribution in Fig. 2 are not observed.

The distribution shown in Fig. 2 is strongly associated with the distribution of various viscous creep features, such as lobate debris aprons, lineated valley fill, concentric crater fill, and terrain softening^{1,5}, that were recognized in Viking images. These viscous creep features have been interpreted to be indicative of ground ice on the basis of thermophysical properties, rheology, and by analogy with features observed in periglacial terrains on Earth^{1,5}. In some of the MOC images, the mapped mantle layer exhibits viscous creep and flow features when present on steep slopes¹² as well as patterned ground (for example, polygons)¹³. Although patterned ground is not uniquely indicative of ground ice, viscous creep requires a rheology strongly consistent with an ice-saturated regolith.

Although there are no direct earth analogues for the terrain shown in Fig. 1, the thermophysical and mechanical properties of frozen loess observed in the Fox permafrost tunnel¹⁴ provide an analogue that explains the primary morphologic characteristics. The tunnel temperature is maintained between -2°C and -5°C and during the winter dry air flowing through the tunnel causes sublimation of ice from the loess, leaving behind a porous but cohesive material. Over the past 30 years a desiccated layer of the order of 10 cm thick has formed on the walls and ceiling. Although this remnant on the walls and ceiling of the tunnel, it collapses to particulate loess upon slight disturbance. The calculated thermal inertia of desiccated loess under Mars conditions indicates that it would not be distinguishable from loose martian dust. A similar, porous, structured material has been produced in experiments designed to model martian polar materials¹⁵.

In general, the surface winds on Mars predicted from models¹⁶ and measured at the landing sites are too weak to disaggregate a mantle with the strength of bonded loess. However, dust devils, the

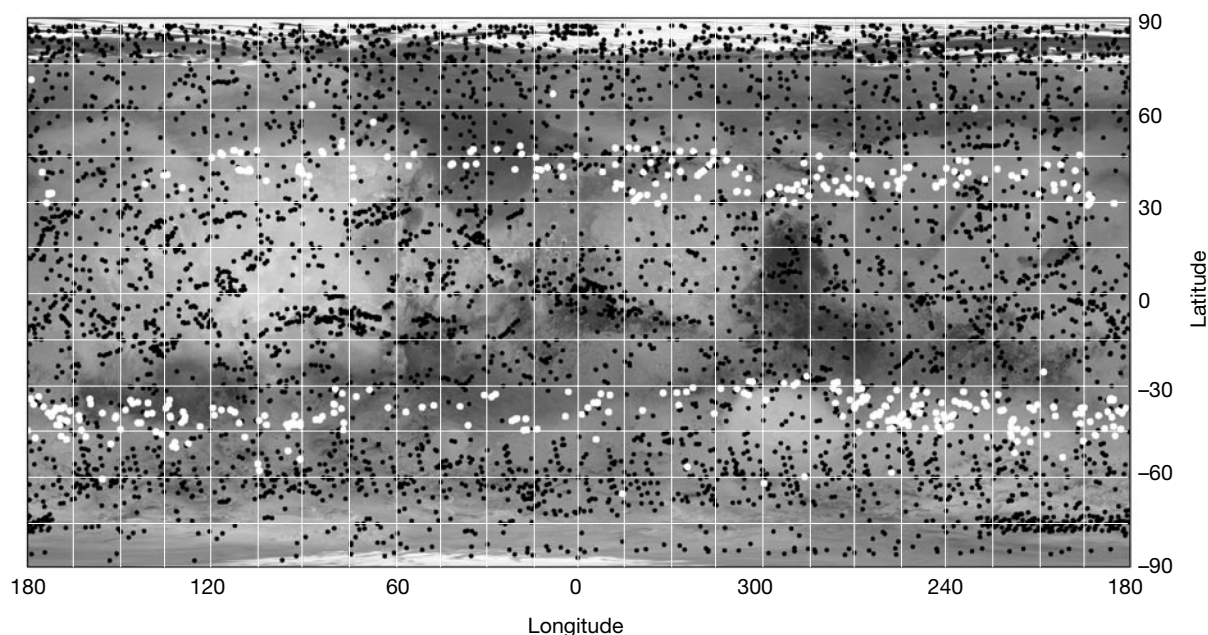


Figure 2 Global distribution of the image locations used in this investigation. Simple cylindrical projection with the global Viking basemap as a background; image locations

that exhibit the dissected terrain are shown as white dots and images examined that do not exhibit such terrain are shown in grey dots.

most common energetic surface component of the martian atmosphere yet characterized, do provide the requisite strengths^{17,18}. High-pressure gradients between the dust-devil core and ambient conditions (up to 50 Pa difference in pressure) coupled with high wind speeds (typically 25–30 m s⁻¹ but may be up to 60 m s⁻¹ in the larger events) create surface shear strengths sufficient to disaggregate the dissected mantle. Material liberated by the passage of the dust devil is lofted and dispersed leaving no accumulations of residual material.

On the basis of this evidence, we conclude that this eroded mantle represents regions on Mars where a layer of generally uniform thickness of dust and soil has been permeated and cemented by ice, but that is now in the process of active disaggregation. Airfall deposition concurrent with ice cementation provides a viable mechanism to produce the uniform thickness and lack of internal structure implied by the morphology. The very smooth character of the intact surfaces implies young surfaces. We have not yet observed an unmodified impact crater in the intact surfaces greater than 100 m in diameter; this indicates a surface age of less than 1.5×10^5 years¹⁹. The bimodal latitude dependence with peaks at

$\pm 40^\circ$ (Fig. 3) is the central defining characteristic of the global distribution.

The striking latitude dependence of this terrain provides confirming evidence for a dynamic ground-ice model for Mars⁷. In this spatially explicit diffusion and condensation model, changes in insolation and atmospheric water abundance driven by variations in the eccentricity, obliquity and time of perihelion (precession) result in the formation and removal of ground ice in the top 2–4 m of soil over timescales of 100,000 years, particularly for latitudes greater than 30° . This model predicts that the most recent maximum extent of soil-hosted ground ice was around 100,000 years ago and reached to within 30° of the equator, but the ice has since retreated. Ground ice is currently predicted to be stable in the martian soil poleward of 40° (refs 4, 7) although the exact latitude depends on the local thermophysical properties. The equatorward limit of the distribution shown in Fig. 3 thus indicates regions of the surface that retain remnant patches of soil-hosted ground ice formed during the last ground-ice maximum, whereas the poleward limit marks where the ice-cemented mantle persists and is stable.

The MOC images and their distribution indicate the following sequence of events for the formation and evolution of this terrain. (1) Under favourable orbital configurations, airfall dust is permeated and cemented by atmospherically deposited ice through a diffusion and condensation process. (2) Changes in insolation driven by the precession of the equinoxes and variations in obliquity lead to surface and soil conditions where ice is no longer stable. (3) Destabilization of the surface (pits, cracks, or sags due to cryospheric processes) allow devolatilization of the near-surface ground ice. (4) The devolatilized dust is analogous to desiccated loess and can be disaggregated and removed by atmospheric processes. (5) Removal of the ice-cemented soil layer reveals the pre-existing surface. With continued orbital variations, this process can repeat itself numerous times over timescales of 10^6 years.

A layer of soil 1–10 m thick with the capacity to store and release water ice on timescales of 10^5 years has important implications for the transport and storage of water on Mars. The 30° – 60° latitude bands are zones of active accumulation and release of ground ice. To estimate the fluxes of water due to this process we use a conservative estimate of 3 m for the active ground ice layer with a porosity range of 10–40%. This results in 30–110 g cm⁻² of water cycled within these latitude bands, which, integrated over the latitude bands, is equivalent to a global layer of water between 10 cm and 40 cm thick or a total volume of $(1.5\text{--}6.0) \times 10^4$ km³. These observations provide direct evidence for a ground ice cycle on Mars operating on Milankovich timescales. This ground ice cycle is the modern manifestation of martian ice ages and has significantly modified the martian surface. □

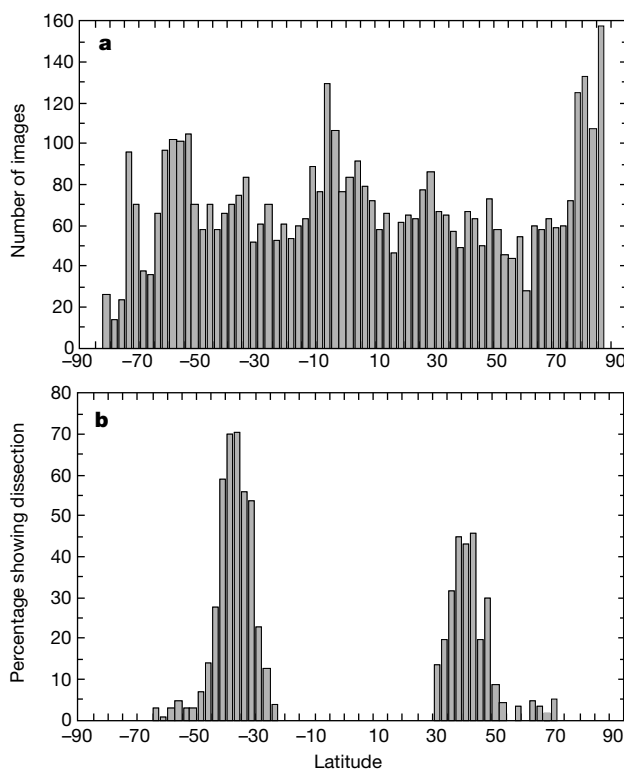


Figure 3 Latitude distribution of MOC frames showing dissection. **a**, Total distribution of images used in this study in 2.5° latitude bins. **b**, Percentage of images shown in **a** that exhibit dissection. For the northern distribution the mean is 41.93° , median is 40.51° , and the standard deviation of the distribution is 10.71° . For the southern distribution the mean is -40.79° , the median -39.95° , and the standard deviation is 10.13° . The equatorward limit of the distributions indicate regions that retain remnant patches of soil-hosted ground ice formed during the last ground-ice maximum. In MOC images that do not exhibit the three components required for identification a uniform mantle is generally absent in the 30° to 40° latitude bands. The poleward limit of the distribution represents regions that are either currently experiencing incipient desiccation and dissection or are remnants of the previous period of minimum ground-ice stability. In MOC images observed for locations poleward of 40° or 50° , a uniform mantle is commonly present. The poleward boundaries of the distributions also correlate with a decrease in topographic roughness calculated with a short 0.6-km baseline using Mars Observer Laser Altimeter (MOLA) data²⁰. Because the decrease in short baseline roughness is not accompanied by changes in long baseline roughness, it is interpreted as due to a uniform mantling of the surface poleward of 50° .

Received 22 December 2000; accepted 5 June 2001.

1. Squyres, S. W. Urey Prize Lecture: Water on Mars. *Icarus* **79**, 229–288 (1989).
2. Carr, M. H. *Water on Mars* (Oxford Univ. Press, New York, 1996).
3. Clifford, S. M. A model for the hydrologic and climatic behavior of water on Mars. *J. Geophys. Res.* **98**, 10973–11016 (1993).
4. Fanale, F. P., Salvail, J. R., Zent, A. P. & Postawko, S. E. Global distribution and migration of subsurface ice on Mars. *Icarus* **67**, 1–18 (1986).
5. Squyres, S. W. & Carr, M. H. Geomorphologic evidence for the distribution of ground ice on Mars. *Science* **231**, 249–252 (1986).
6. Houben, W., Haberle, R. M., Young, R. E. & Zent, A. P. Modeling the martian seasonal water cycle. *J. Geophys. Res.* **102**, 9069–9094 (1997).
7. Mellon, M. T. & Jakosky, B. M. The distribution and behavior of Martian ground ice during past and present epochs. *J. Geophys. Res.* **100**, 11781–11799 (1995).
8. Malin, M. C. et al. Early view of Martian surface from the Mars Orbiter Camera of Mars Global Surveyor. *Science* **279**, 1681–1685 (1998).
9. Edgett, K. S. & Malin, M. C. New views of Mars eolian activity, materials, and surface properties: Three vignettes from the Mars Global Surveyor Mars Orbiter Camera. *J. Geophys. Res.* **105**, 1623–1650 (2000).
10. Christensen, P. R. & Moore, H. J. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 686–729 (Univ. Arizona Press, Tucson, 1992).
11. Cooper, C. D. & Mustard, J. F. Criteria for remote sensing detection of sulfate cemented soils on Mars. *Lunar Planet. Sci.* [online] **31**, 1747 (2000).
12. Mustard, J. F. Recent ground ice on Mars and implications for volatiles and surface processes. *Lunar Planet. Sci.* [online] **32**, 1988 (2001).

13. Seibert, N. M. & Kargel, J. S. Small-scale martian polygonal terrain: Implications for liquid surface water. *Geophys. Res. Lett.* **28**, 899–902 (2001).
14. Johnson, J. J. & Lorenz, R. D. Thermophysical properties of Alaskan loess: An analog material for the martian polar layered terrain? *Geophys. Res. Lett.* **27**, 2769–2772 (2000).
15. Storrs, A. D., Fanale, F. P., Saunders, R. S. & Stephens, J. B. The formation of filamentary sublimite residues (FSR) from mineral grains. *Icarus* **76**, 493–512 (1988).
16. Haberle, R. M., Houben, W., Barnes, J. R. & Young, R. E. A simplified three-dimensional model for Martian climate studies. *J. Geophys. Res.* **102**, 9051–9068 (1997).
17. Metzger, S. M. *et al.* Dust devil vortices seen by the Mars Pathfinder camera. *Geophys. Res. Lett.* **26**, 2781–2784 (1999).
18. Renno, N. O. *et al.* Martian and terrestrial dust devils: Test of a scaling theory using Pathfinder data. *J. Geophys. Res.* **105**, 1859–1865 (2000).
19. Herkenhoff, K. E. & Plaut, J. J. Surface ages and resurfacing rates of the polar layered deposits on Mars. *Icarus* **144**, 243–253 (2000).
20. Kreslavsky, M. A. & Head, J. W. Kilometer scale roughness of Mars: Results from MOLA data analysis. *J. Geophys. Res.* **105**, 26695–26712 (2000).

Acknowledgements

Discussions were provided by J. Head and P. Schultz and formal reviews were provided by M. Carr and B. Jakosky. The efforts of the MOC science team in creating the data base is gratefully acknowledged. This work is supported by grants from NASA.

Correspondence and requests for materials should be addressed to J.F.M.
(e-mail: john_mustard@brown.edu).

In situ detection of collisionless reconnection in the Earth's magnetotail

M. Øieroset*, T. D. Phan*, M. Fujimoto†, R. P. Lin* & R. P. Lepping‡

* Space Sciences Laboratory, University of California, Berkeley, California 94720, USA

† Department of Earth and Planetary Sciences, Tokyo Institute of Technology, Meguro 152, Japan

‡ NASA Goddard Space Flight Center, Code 696, Greenbelt, Maryland 20771, USA

Magnetic reconnection is the process by which magnetic field lines of opposite polarity reconfigure to a lower-energy state, with the release of magnetic energy to the surroundings. Reconnection at the Earth's dayside magnetopause and in the magnetotail allows the solar wind into the magnetosphere^{1,2}. It begins in a small 'diffusion region', where a kink in the newly reconnected lines produces jets of plasma away from the region. Although plasma jets from reconnection have previously been reported^{3–7}, the physical processes that underlie jet formation have remained poorly understood because of the scarcity of *in situ* observations of the minuscule diffusion region. Theoretically, both resistive and collisionless processes can initiate reconnection^{8–14}, but which process dominates in the magnetosphere is still debated. Here we report the serendipitous encounter of the Wind spacecraft with an active reconnection diffusion region, in which are detected key processes predicted by models^{8–13} of collisionless reconnection. The data therefore demonstrate that collisionless reconnection occurs in the magnetotail.

When magnetic field lines of opposite polarity convect toward each other they reconnect and change configuration in a small diffusion region centred around a magnetic X-line (Fig. 1). Outside the diffusion region, in the inflow as well as the high-speed jet regions, the magnetic field lines are 'frozen' to the plasma. In the diffusion region, however, the field lines diffuse from the plasma, allowing the opposite field lines to merge and change partners (Fig. 1b). The occurrence of reconnection depends critically on the plasma processes within the diffusion region. Essentially all of the experimental tests for its occurrence, however, pertain to

observations outside the diffusion region. Much of our knowledge of processes in the diffusion region thus derives solely from theoretical modelling^{8–15}.

According to theory, reconnection can be accomplished either by resistive or by collisionless mechanisms. The classical collision rate in the magnetosphere is low, but anomalous resistivity could be provided by increased wave activity^{15–17}. If the resistive scale size is larger than the ion scale size (the ion skin depth—see Fig. 1 for definition), reconnection is accomplished by resistive diffusion of ions and electrons at the resistive scale and no separation between ions and electrons occurs. If, on the other hand, the resistive scale size is smaller than the ion skin depth, collisionless effects become dominant, and a separation of ions and electrons (the Hall effect) occurs as the ions are first unfrozen from the magnetic field (at the larger ion scale) while the electrons continue to carry the field lines toward the X-line and eventually diffuse at the smaller electron scale¹⁴. The separation between electrons and ions in the ion diffusion region gives rise to a system of currents, termed the 'Hall current' (see Fig. 1b), which in turn induces a quadrupolar out-of-plane Hall magnetic field pattern⁸. The Hall quadrupolar fields should be recognizable by a spacecraft traversing the magnetotail reconnection region, from one side of the X-line to the other, as a reversal of the out-of-plane component (*y*-component) of the magnetic field, with an amplitude of ~30% of the background magnetospheric field¹⁸. Electrons carrying the Hall current are expected along the boundary separating the lobe and the plasma sheet^{19–21}.

Out-of-plane magnetic fields have recently been detected by the Geotail spacecraft in the high-speed flow region outside the diffusion region both in the near-Earth tail²¹ and at the dayside magnetopause²². These fields are presumably the extensions of the Hall magnetic fields originating inside the diffusion region. But as long as such signatures are obtained only outside the diffusion region one cannot tell with certainty whether the observed signatures are the cause or consequence of reconnection. Only *in situ* detection of Hall signatures in the diffusion region would unambiguously point to the dominance of collisionless effects in magnetic reconnection.

The Wind spacecraft made a fortuitous direct encounter with the ion diffusion region on 1 April 1999. As the spacecraft travelled away from the Sun through the Earth's magnetotail (see Fig. 1), at about 60 Earth radii behind the Earth it detected a period of Earthward-directed plasma jets followed by an interval of tailward-directed jets, with proton flow speed reaching 400 km s⁻¹ (Fig. 2b). The velocities of both the Earthward and tailward jets are in agreement with the predicted slingshot effect resulting from reconnection⁷. This flow interval was embedded in a 10-hour interval of quasi-steady reconnection⁷. The detection of the reversal of the reconnection jets implies that the spacecraft moved from the Earthward side to the tailward side of an active reconnection X-line (Fig. 1b). The spacecraft remained in the reconnection layer as it crossed from one side of the X-line to the other, as evident from the uninterrupted transition from Earthward to tailward flow (Fig. 2a), so the spacecraft must have crossed the diffusion region. If the crossing from one side of the X-line to the other had occurred outside the diffusion region, the spacecraft would have detected an interval of slow flow in the *z*-direction only, which corresponds to the reconnection inflow region⁷. It can be inferred from the mostly negative sign of the *x*-component of the magnetic field, *B_x*, that the spacecraft stayed predominantly on the southward side of the plasma sheet midplane (the neutral sheet) during the crossing near the X-line, as schematically shown in Fig. 1b.

We now describe the Hall electron and magnetic field signatures obtained in the diffusion region implying that collisionless (rather than resistive) effects dominate in this event.

As the spacecraft travelled from the region of Earthward jets to the region of tailward jets while remaining mostly below the plasma

sheet midplane, it observed a decrease followed by an increase in y -component of the magnetic field, B_y . The amplitudes of these changes were about 4.5 nT superposed on a background B_y field of 6 nT. The sense of the field reversal for this spacecraft trajectory is in precise agreement with the predicted quadrupolar out-of-plane Hall magnetic field configuration^{8,23} (Fig. 1b). Even more convincing, shortly after the flow reversal Wind twice crossed briefly to the northward side of the neutral sheet, as apparent from the positive sign of the B_x component, before returning to the southward side. During the two brief excursions into the northern hemisphere the B_y component actually turned negative, as required by the model shown in Fig. 1b. The observed Hall magnetic field amplitude of 4.5 nT corresponds to about 40% of the total magnetic field magnitude (about 12 nT), close to the prediction based on a two-dimensional hybrid simulation¹⁸. The presence of nearly symmetric bipolar B_y field superposed on a background B_y field of 6 nT (50% of the total field) is consistent with theoretical predictions that the presence of a guide field does not change the quadrupolar Hall field pattern in a significant way^{12,24}.

In addition to the Hall magnetic field signatures, a low-energy (<300 eV) electron beam aligned with the magnetic field and directed towards (implying a current away from) the X-line was observed just before the flow reversal (Fig. 3). The electron beam was observed only for a 2-minute interval preceding the flow reversal when the B_x component was large and negative, which indicates that the spacecraft was near the boundary between the plasma sheet and the lobe in the southern hemisphere. The sense of the electron beam and the confinement of this beam to the plasma sheet/lobe boundary are consistent with these electrons being the Hall current carriers (see Fig. 1b).

A dip in the density was detected just preceding the time of the flow reversal (Fig. 2a). A density depletion along the separatrices⁹ and at the centre of the diffusion region^{10,12} has been predicted in recent simulation models. The observed density depletion (Fig. 2a) is not centred at the time of the flow reversal and would correspond more to a dip along the separatrices, although a brief entry into the lobe is also a possible interpretation.

The length of the ion diffusion region (along the x -axis) is difficult to estimate from single spacecraft observations. If, however, one assumes the X-line to be stationary in space while the spacecraft moves at a speed of about 1 km s⁻¹ in the negative x -direction one obtains a diffusion region length of 1,300 km, or roughly 2 ion skin depths, for a crossing duration of 21 minutes. The true length is larger or smaller depending on whether the X-line moved with or against the spacecraft trajectory.

Waves with power between the lower hybrid resonance and the ion cyclotron frequency were observed (not shown) throughout the many hours of observed high-speed jets. The magnetic spectral energy density was similar inside and outside the diffusion region. Although these waves could be interpreted as whistlers generated in the collisionless diffusion region⁹, we cannot rule out that the waves could be locally generated by processes outside the diffusion region.

The observations presented here constitute a rare encounter with an active reconnection diffusion region in the Earth's magnetotail where Hall effects indicative of collisionless reconnection were detected. The event was embedded in a 10-hour interval of quasi-steady reconnection. Evidence for Hall effects has also been reported outside the diffusion region in highly time-varying reconnection closer to the Earth^{19,21} and at the dayside magnetopause²², and perhaps in the diffusion region in the polar cusp (J. D. Scudder,

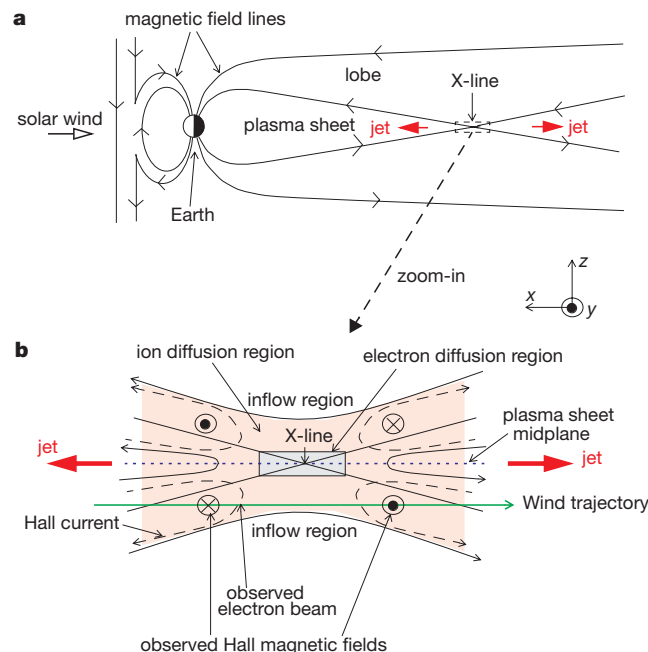


Figure 1 Schematics of magnetic reconnection in the Earth's magnetosphere. **a**, The noon-midnight plane of the magnetosphere showing dayside magnetopause and magnetotail reconnection. In the magnetotail, oppositely directed magnetic field lines in the northern and southern lobe regions convect towards the low-latitude plasma sheet, touch each other and reconnect at an X-point or X-line (in the out-of-plane y -direction). In the process magnetic energy is converted into kinetic energy in the form of bi-directional plasma jets directed towards and away from the Earth. **b**, Zoom-in on the region surrounding the X-line. In the collisionless regime, oppositely directed field lines convect towards each other from the top and the bottom of the figure and reconnect at a distance of the order of the ion skin depth (defined as $m/(ne^2\mu_0)$ where m is the mass of the species, n is the plasma density, e is the charge, and μ_0 is the permeability in vacuum)

from the X-line. The skin depth in the plasma sheet is about 700 km for ions and 20 km for electrons. The ion diffusion region is marked by the shaded red area, and the smaller electron diffusion region is shown as a grey box. In the ion diffusion region, the separation between ions and electrons at the ion scale creates a system of Hall currents. These in turn induce a quadrupolar out-of-plane magnetic field pattern⁸ which was observed by the Wind spacecraft as it travelled from the Earthward to the tailward side of the X-line. Electron motion consistent with the Hall current was also detected near the boundary between lobe and plasma sheet as expected from the schematic. The coordinate system is defined such that x points towards the Sun, z is normal to the current sheet, and y is directed out of the plane.

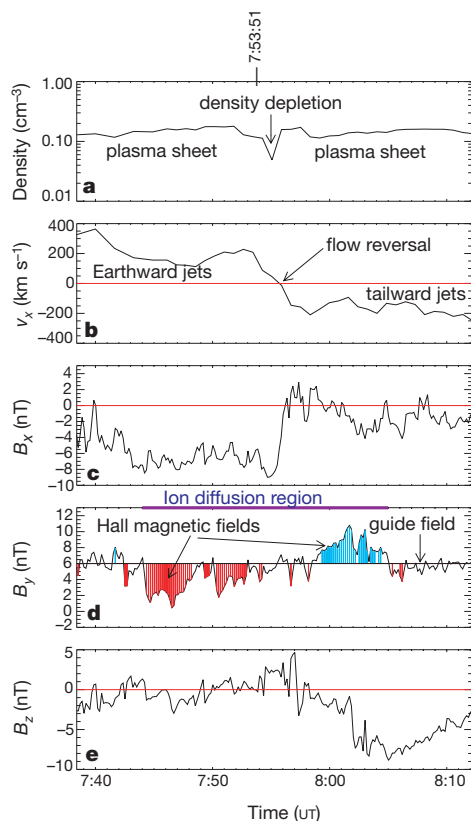


Figure 2 Observations by Wind in the ion diffusion region. The ion diffusion region (marked) is identified from the B_y component as the region of large 'out-of-plane' Hall magnetic fields surrounding the flow reversal. **a**, The plasma density, which was at the typical plasma sheet values throughout the interval except for a brief (~1 minute) dip to lower density just before the flow reversal. **b**, The x -component of the proton flow velocity, showing bi-directional jets and a flow reversal. **c–e**, The three components of the magnetic field. The direction normal to the current sheet was determined by minimum variance analysis²⁸ of the magnetic field data. Bipolar B_y field variations with polarities consistent with the Hall magnetic field pattern (Fig. 1b) were detected as the spacecraft crossed from the Earthward side to the tailward side of the flow reversal region. To highlight the bipolar B_y signature, magnetic field deviations (from the guide field) larger than 1.5 nT are shaded blue for negative changes and red for positive enhancements.

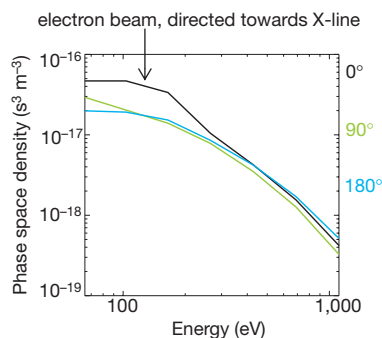


Figure 3 Detection by Wind of a low-energy electron beam carrying the Hall current. The electron distributions at 0°, 90° and 180° to the magnetic field were obtained at 7:53:51 Universal Time (UT), just before the flow reversal when the spacecraft was in the southern hemisphere (below the plasma sheet midplane) near the boundary between the lobe and the plasma sheet. The electron distribution is isotropic at all energies, except for the appearance of a field-aligned electron beam with energy up to 300 eV. The detection of the field-aligned (0°) electron beam implies that the beam was directed towards the X-line, consistent with the prediction that the Hall current should be directed away from the X-line at this location (see Fig. 1b).

personal communication). Together these observations indicate that magnetic reconnection in the entire magnetosphere operates in the collisionless regime at least some, if not most, of the time.

According to theory^{9–13}, reconnection proceeds at a much faster rate when it is mediated by collisionless rather than resistive effects, with predicted (dimensionless) collisionless reconnection rates reaching ~0.2 (refs 9–13). A collisionless reconnection regime in the Earth's magnetosphere thus implies much higher rates of solar wind entry than depicted by present-day global numerical simulation models of interactions between the solar wind and magnetosphere, which assume resistive reconnection^{25–27}.

The detection of Hall effects in the ion diffusion region implies the importance of collisionless (non-resistive) effects in reconnection. But the processes that ultimately cause the electrons to diffuse from the magnetic field in the much smaller electron diffusion region remain one of the main unsolved problems in reconnection research. The identification of electron processes requires high-resolution measurements capable of resolving electron-scale structures and dynamics. Such observations could then be used to verify a recent rather surprising theoretical finding, based on two-dimensional reconnection configurations, that ion-scale Hall effects ultimately determine the reconnection rate, irrespective of which electron-scale processes eventually cause the electrons to diffuse from the magnetic field^{9–13}. □

Received 1 May; accepted 12 June 2001.

- Dungey, J. W. Interplanetary magnetic field and the auroral zones. *Phys. Rev. Lett.* **6**, 47–48 (1961).
- Cowley, S. W. H. In *Magnetic Reconnection in Space and Laboratory Plasmas* (ed. Hones, E. W.) 375–378 (Geophysics Monograph 30, American Geophysical Union, Washington DC, 1984).
- Paschmann, G. *et al.* Plasma acceleration at the Earth's magnetopause: Evidence for reconnection. *Nature* **282**, 243–246 (1979).
- Sonnerup, B. U. Ö. *et al.* Evidence for reconnection at the earth's magnetopause. *J. Geophys. Res.* **86**, 10049–10067 (1981).
- Gosling, J. T., Thomsen, M. F., Bame, S. J. & Russell, C. T. Accelerated plasma flows at the near-tail magnetopause. *J. Geophys. Res.* **91**, 3029–3041 (1986).
- Phan, T. D. *et al.* Extended magnetic reconnection at the earth's magnetopause from detection of bi-directional jets. *Nature* **404**, 848–850 (2000).
- Øieroset, M., Phan, T. D., Lin, R. P. & Sonnerup, B. U. Ö. Walén and variance analyses of high-speed flows observed by Wind in the midtail plasma sheet: Evidence for reconnection. *J. Geophys. Res.* **105**, 25247–25263 (2000).
- Sonnerup, B. U. Ö. In *Solar System Plasma Physics*, Vol. III (eds Lanzerotti, L. T., Kennel, C. F. & Parker, E. N.) 45–108 (North-Holland, New York, 1979).
- Shay, M. A., Drake, J. F., Rogers, B. N. & Denton, R. E. Alfvénic collisionless magnetic reconnection and the Hall term. *J. Geophys. Res.* **106**, 3759–3772 (2001).
- Ma, Z. W. & Bhattacharjee, A. Hall magnetohydrodynamic reconnection: The Geospace Environment Modeling challenge. *J. Geophys. Res.* **106**, 3773–3782 (2001).
- Hesse, M., Birn, J. & Kuznetsova, M. Collisionless magnetic reconnection: Electron processes and transport modeling. *J. Geophys. Res.* **106**, 3721–3735 (2001).
- Pritchett, P. L. Geospace Environment Modeling magnetic reconnection challenge: Simulations with a full particle electromagnetic code. *J. Geophys. Res.* **106**, 3783–3798 (2001).
- Birn, J. *et al.* Geospace Environmental Modeling (GEM) magnetic reconnection challenge. *J. Geophys. Res.* **106**, 3715–3719 (2001).
- Vasyliunas, V. M. Theoretical models of magnetic field line merging. I. *Rev. Geophys. Space Phys.* **13**, 303–336 (1975).
- Coroniti, F. V. & Eviatar, A. Magnetic field reconnection in collisionless plasma. *Astrophys. Suppl.* **33**, 189–210 (1977).
- Cattell, C. *et al.* ISEE-1 and Geotail observations of low frequency waves at the magnetopause. *J. Geophys. Res.* **100**, 11823–11829 (1995).
- Sibeck, D. G. *et al.* In *Magnetospheric Plasma Sources and Losses* Ch. 5 (eds Hultqvist, B. *et al.*) 207–283 (Space Sciences Series of ISSI 6, Kluwer Academic, Dordrecht, 1999).
- Shay, M. A., Drake, J. F., Rogers, B. N. & Denton, R. E. The scaling of collisionless, magnetic reconnection for large systems. *Geophys. Res. Lett.* **26**, 2163–2166 (1999).
- Fujimoto, M. *et al.* Observations of earthward streaming electrons at the trailing boundary of a plasmoid. *Geophys. Res. Lett.* **24**, 2893–2896 (1997).
- Hoshino, M., Mukai, T., Yamamoto, T. & Kokubun, S. Ion dynamics in magnetic reconnection: Comparison between numerical simulations and Geotail observations. *J. Geophys. Res.* **103**, 4509–4530 (1998).
- Nagai, T. *et al.* Geotail observations of the Hall current system. Evidence of magnetic reconnection in the magnetotail. *J. Geophys. Res.* (in the press).
- Deng, X. H. & Matsumoto, H. Rapid magnetic reconnection in the Earth's magnetosphere generated by whistler waves. *Nature* **410**, 557–559 (2001).
- Terasawa, T. Hall current effect on tearing mode instability. *Geophys. Res. Lett.* **10**, 475–478 (1983).
- Wang, X., Bhattacharjee, A. & Ma, Z. W. Collisionless reconnection: Effects of Hall current and electron pressure gradient. *J. Geophys. Res.* **105**, 27633–27648 (2000).
- Goodrich, C. C., Lyon, J. G., Wiltberger, M., Lopez, R. E. & Papadopoulos, K. An overview of the impact of the January 10–11, 1997, magnetic cloud on the magnetosphere via global MHD simulations. *Geophys. Res. Lett.* **25**, 2537–2540 (1998).

26. Raeder, J. Modeling the magnetosphere for northward interplanetary magnetic field: Effects of electrical resistivity. *J. Geophys. Res.* **104**, 17357–17368 (1999).
27. Gombosi, T. I., DeZeeuw, D. L., Groth, C. P. T., Powell, K. G. & Song, P. in *Physics Space Plasmas* Vol. 15 (eds Chang, T. & Jasperse, J. R.) 121–128 (MIT Press, Cambridge, Massachusetts, 1998).
28. Sonnerup, B. U. Ö. & Cahill, L. J. Magnetopause structure and attitude from Explorer 12 observations. *J. Geophys. Res.* **72**, 171–183 (1967).

Acknowledgements

We thank S. Bale, C. Chaston, B. Sonnerup, J. Drake and A. Bhattacharjee for discussions and T. Nagai for providing a preprint of his Geotail manuscript at an early stage. This work is supported by grants from NASA.

Correspondence and requests for materials should be addressed to M.Ø.
(e-mail: oieroset@ssl.berkeley.edu).

Quantum-enhanced positioning and clock synchronization

Vittorio Giovannetti*, Seth Lloyd† & Lorenzo Maccone*

* Massachusetts Institute of Technology, Research Laboratory of Electronics,
MIT36-497, Cambridge, Massachusetts 02139, USA

† Massachusetts Institute of Technology, Department of Mechanical Engineering,
MIT3-160, Cambridge, Massachusetts 02139, USA

A wide variety of positioning and ranging procedures are based on repeatedly sending electromagnetic pulses through space and measuring their time of arrival. The accuracy of such procedures is classically limited by the available power and bandwidth. Quantum entanglement and squeezing have been exploited in the context of interferometry^{1–3}, frequency measurements⁶, lithography⁷ and algorithms⁸. Here we report that quantum entanglement and squeezing can also be employed to overcome the classical limits in procedures such as positioning systems, clock synchronization and ranging. Our use of frequency-entangled pulses to construct quantum versions of these protocols results in enhanced accuracy compared with their classical analogues. We describe in detail the problem of establishing a position with respect to a fixed array of reference points.

Any position (say, that of Alice) may be obtained simply by sending pulses that originate from that position and measuring the time it takes for each pulse to reach the reference points. The time of flight, the speed of the pulses and the arrangement of the reference points determine Alice's position. The accuracy of such a procedure depends on the number of pulses, their bandwidth and the number of photons per pulse. Here we show that, by measuring the correlations between the times of arrival of M pulses which are frequency-entangled, it is in principle possible to increase the accuracy of such a positioning procedure by a factor $\sqrt{M}$ as compared to positioning using unentangled pulses with the same bandwidth. Moreover, if number-squeezed pulses can be produced⁹, it is possible to obtain a further increase in accuracy of $\sqrt{N}$ by employing squeezed pulses of N quanta, rather than employing 'classical' coherent states with a mean number of quanta N . Combining entanglement with squeezing gives an overall enhancement of $\sqrt{MN}$. In addition, the procedure exhibits improved security: because the timing information resides in the entanglement between pulses, it is possible to implement quantum cryptographic schemes that do not allow an eavesdropper to obtain information on the position of Alice (V.G., S.L. and L.M., unpublished results). The primary drawbacks of this scheme are the difficulty of creating the requisite entanglement and the sensitivity to loss. On the other hand, the frequency entanglement allows similar schemes to be highly robust against pulse broadening due to transit through dispersive media¹⁰.

The clock synchronization problem can be treated using the same method. In refs 11 and 12 two techniques for clock synchronization using entangled states were presented. However, the authors of ref. 11 themselves point out that the resources needed by their scheme could be used to perform conventional clock synchronization without entanglement. Similarly, all the enhancement reported in ref. 12 arises from employing high-frequency atoms, which themselves could be used for clock synchronization to the same degree of accuracy without any entanglement. In neither case do these proposals give an obvious enhancement over classical procedures that use the same resources. Here, by contrast, we show that quantum features such as entanglement and squeezing could in principle be used to supply a significant enhancement of the accuracy of clock synchronization as compared with classical protocols using light of the same frequency and power. In fact, the clock synchronization could be accomplished by sending pulses back and forth between the parties whose clocks are to be synchronized, and measuring the times of arrival of the pulses (Einstein's protocol). In this way synchronization may be treated on the same basis as positioning, and the same accuracy enhancements may be achieved through entanglement and squeezing. Here we will address in detail only the enhancement of positioning accuracy.

In order to introduce the formalism, we present the simple case of position measurement with classical coherent pulses. As each dimension can be treated independently, the analysis will be limited to the one-dimensional case. For the sake of simplicity, consider the situation in which Alice wants to measure her position x by sending a pulse to each of M detectors placed in a known position (Fig. 1). This can be easily generalized to different set-ups, such as the case in which the detectors are not all in the same location, the case in which only one detector is employed with M time-separated pulses, the case in which the pulses originate from the reference points and are measured by Alice (as in GPS, the global positioning system), and so on. Alice's estimate of her position is given by $x = (c/M)\sum_{i=1}^M t_i$, where t_i is the travel time of the i th pulse and c is the speed of light. The variable t_i has an intrinsic indeterminacy dependent on the spectral characteristics and mean number of photons N of the i th pulse. For example, given a gaussian pulse of frequency spread $\Delta\omega$, then, according to the central limit theorem, t_i cannot be measured with an accuracy better than $1/(\Delta\omega\sqrt{N})$ as it is estimated at most from N data points (that is, the times of arrival of the single photons, each having an indeterminacy $1/\Delta\omega$). Thus, if Alice uses M gaussian pulses of equal frequency spread, the accuracy in the measurement of the average time of arrival is:

$$\Delta t = \frac{1}{\Delta\omega\sqrt{MN}} \quad (1)$$

Quantum mechanics allows us to do much better. In order to demonstrate the gain in accuracy afforded by quantum mechanics, we provide first a fully quantum analysis of the problem of determining the average time of arrival of a set of M classical pulses, each having a mean number of photons N . Such a quantum treatment for a classical problem may seem excessive, but once the quantum formalism is presented, the speed-up attainable in the fully quantum case can be derived directly. In addition, it is important to verify that no improvement over equation (1) is obtainable using classical pulses. The M coherent pulses are described by a state of the radiation field of the form

$$|\Psi\rangle_{\text{cl}} \equiv \bigotimes_{i=1}^M \bigotimes_{\omega} \alpha \left(\phi_{\omega} \sqrt{N} \right) \quad (2)$$

where ϕ_{ω} is the pulses' spectral function, $|\alpha(\lambda_{\omega})\rangle_i$ is a coherent state of amplitude λ_{ω} in the mode at frequency ω directed towards the i th detector, and N is the mean number of photons in each pulse. The pulse spectrum $|\phi_{\omega}|^2$ has been normalized so that $\int d\omega |\phi_{\omega}|^2 = 1$. For detectors with perfect time resolution, the joint probability for the

i th detector to detect N_i photons in the i th pulse at times $t_{i,k}$ is given by¹³

$$p(\{t_{i,k}\}) \propto \left\langle : \prod_{i=1}^M \prod_{k=1}^{N_i} E_i^{(-)}(t_{i,k}) E_i^{(+)}(t_{i,k}) : \right\rangle \quad (3)$$

where $t_{i,k}$ is the time of arrival of the k th photon in the i th pulse, shifted by the position of the detectors $t_{i,k} \rightarrow t_{i,k} + x/c$. The signal field at the position of the i th detector at time t is given by $E_i^{(-)}(t) \equiv \int d\omega a_i^\dagger(\omega) e^{i\omega t}$ and $E_i^{(+)} \equiv (E_i^{(-)})^\dagger$, where $a_i(\omega)$ is the field annihilator of a quantum of frequency ω at the i th detector, which satisfies $[a_i(\omega), a_j^\dagger(\omega')] = \delta_{ij} \delta(\omega - \omega')$. The estimation of the ensemble average in equation (3) on the state $|\Psi\rangle_{\text{cl}}$, using the property $a(\omega') \otimes |\alpha(\lambda_\omega)\rangle = \lambda_\omega \otimes |\alpha(\lambda_\omega)\rangle$, gives

$$p(\{t_{i,k}\}) \propto \prod_{i=1}^M \prod_{k=1}^{N_i} |g(t_{i,k})|^2 \quad (4)$$

where $g(t)$ is the Fourier transform of the spectral function ϕ_ω . Averaging over the times of arrival $t_{i,k}$ and over the number of photons N_i detected in each pulse, we have:

$$\langle t \rangle = \left\langle \frac{1}{M} \sum_{i=1}^M \frac{1}{N_i} \sum_{k=1}^{N_i} t_{i,k} \right\rangle = \bar{\tau}; \Delta t \geq \frac{\Delta\tau}{\sqrt{MN}} \quad (5)$$

with approximate equality for $N \gg 1$. Here $\bar{\tau} \equiv \int dt |g(t)|^2 t$ and $\Delta\tau^2 \equiv \int dt |g(t)|^2 (t - \bar{\tau})^2$ are independent of i and k as all the photons have the same spectrum. Equation (5) is the generalization of (1) for non-gaussian pulses.

Quantum light can exhibit phenomena that are not possible classically, such as entanglement and squeezing, which, as will now be seen, can give significant enhancement for determining the average time of arrival. We first consider entanglement. The framework just established allows the direct comparison between frequency-entangled pulses and unentangled ones. For the sake of clarity, we consider initially single-photon entangled pulses.

We define the 'frequency state' $|\omega\rangle$ for the electromagnetic field as the state in which all modes are in the vacuum state, except for the mode at frequency ω which is populated by one photon. Thus the state $\int d\omega \phi_\omega |\omega\rangle$ represents a single-photon wave packet with spectrum $|\phi_\omega|^2$. We consider the M -photon frequency-entangled state given by

$$|\Psi\rangle_{\text{en}} \equiv \int d\omega \phi_\omega |\omega\rangle_1 |\omega\rangle_2 \cdots |\omega\rangle_M \quad (6)$$

where the subscripts 1, 2, M indicate the detector each photon is travelling to. Inserting $|\Psi\rangle_{\text{en}}$ in equation (3), and specializing to the case $N_i = 1$, it follows that

$$p(t_1, \dots, t_M) \propto \left| g\left(\sum_{i=1}^M t_i\right) \right|^2 \quad (7)$$

That is, the entanglement in frequency translates into the bunching of the times of arrival of the photons of different pulses: although their individual times of arrival are random, the average $t \equiv (1/M) \sum_{i=1}^M t_i$ of these times is highly peaked. (The measurement of t follows from the correlations in the times of arrival at the different detectors.) Indeed, from equation (7) it turns out that the probability distribution of t is $|g(Mt)|^2$. This immediately implies that the average time of arrival is determined to an accuracy

$$\Delta t = \frac{\Delta\tau}{M} \quad (8)$$

where $\Delta\tau$ is the same as in equation (5). This result shows a $\sqrt{M}$ improvement over the classical case of equation (5).

To emphasize the importance of entanglement, equation (8) should be compared to the result that would be obtained from an unentangled state analogous to $|\Psi\rangle_{\text{en}}$. To this end, we consider the state defined as

$$|\Psi\rangle_{\text{un}} \equiv \bigotimes_{i=1}^M \int d\omega_i \phi_{\omega_i} |\omega_i\rangle_i \quad (9)$$

which describes M uncorrelated single-photon pulses each with spectral function ϕ_ω . By looking at the spectrum of the state obtained by tracing away all but one of the modes in equation (6), each of the photons in equation (9) can be shown to have the same spectral characteristics as the photons in the entangled state $|\Psi\rangle_{\text{en}}$. Now, using equation (3) for the uncorrelated M photon pulses $|\Psi\rangle_{\text{un}}$, it follows that

$$p(t_1, \dots, t_M) \propto \prod_{i=1}^M |g(t_i)|^2 \quad (10)$$

which is the same result that was obtained for the classical state $|\Psi\rangle_{\text{cl}}$. Thus equation (5) holds, with $N = 1$, also for $|\Psi\rangle_{\text{un}}$. From the comparison of equation (5) and (8), we see that employing frequency-entangled pulses gives an increase in accuracy by a factor $\sqrt{M}$ in the measurement of t with respect to the case of unentangled photons.

As $|\Psi\rangle_{\text{en}}$ is tailored to give the least indetermination in the quantity t , it is appropriate for the geometry of the case given in Fig. 1, where the sum of the times of arrival is needed. Other entangled states can be tailored for different geometric dispositions of the detectors, as will be shown through some examples.

How can we create the needed entangled states? In the case $M = 2$, the twin beam state at the output of a continuous-wave pumped parametric down-converter will be shown to be appropriate. It is a two-photon frequency-entangled state of the form $\int d\omega \phi_\omega |\omega\rangle_s |\omega_0 - \omega\rangle_i$, where ω_0 is the pump frequency and s and i refer to the signal and idler modes respectively. This state is similar to $|\Psi\rangle_{\text{en}}$, and can be employed for position measurements when the two reference points are in opposite directions—for example, one to the left and one to the right of Alice. In fact, it can be seen that

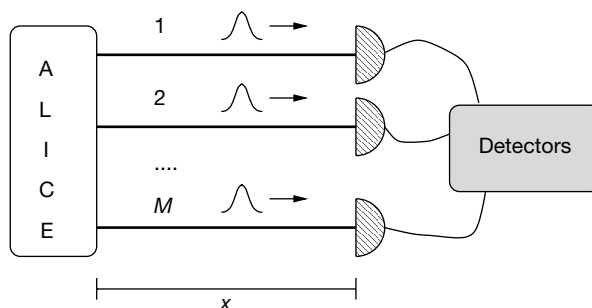


Figure 1 Sketch of the idealized experimental configuration. Alice sends M light pulses to the M detectors. She averages the times of arrival t_i of the pulses to recover her unknown position x .

$p(t_1, t_2) \propto |g(t_1 - t_2)|^2$, and hence such a state is optimized for measurements of differences in the times of arrival, as experimentally reported in ref. 14. In the case of $M = 3$, a suitable state can be obtained starting from a three-photon generation process that creates a state of the form $\int d\omega d\omega' f(\omega, \omega') |\omega\rangle |\omega'\rangle |\omega_0 - \omega - \omega'\rangle$, and then performing a non-demolition (or a post-selection) measurement of the frequency difference of two of the photons. This would create a maximally entangled three-photon state, tailored for the case in which Alice has one detector on one side and two detectors on the other side. However, for $M > 2$, the creation of such frequency-entangled states represents a continuous variable generalization of the Greenberger–Horne–Zeilinger state, and, as such, is a considerable experimental challenge.

Now we turn to the use of number-squeezed states to enhance positioning. Quantum effects in the propagation of multi-photon states are well known (see, for example, ref. 15). The N th excitation of a quantum system (that is, the state $|N\rangle$ of exactly N quanta) has a de Broglie frequency N times the fundamental frequency of the state. Its shorter wavelength makes such a state appealing for positioning protocols. In this case, the needed ‘frequency state’ is given by $|N_\omega\rangle$, defined as the state where all modes are in the vacuum except for the mode at frequency ω , which is in the Fock state $|N\rangle$. The probability of measurement of N quanta in a single pulse at times $t_1, \dots, t_N$ is given by equation (3) with $M = 1$ detectors. We see that, for a state of the form $\int d\omega \phi_\omega |N_\omega\rangle$, the time of arrival probability is given by:

$$p(t_1, \dots, t_N) \propto \left| g \left(\sum_{k=1}^N t_k \right) \right|^2 \quad (11)$$

Such a result must be compared to what one would obtain employing a classical pulse $|\Psi\rangle_{cl}$ of mean number of photons N , that is, the state equation (2) with $M = 1$. Its probability (equation (4)) shows that employing the N -photon Fock state gives an accuracy increase of $\sqrt{N}$ compared to the coherent state with mean number of photons N . The similarity of this result (equation (11)) to that obtained in equation (7) stems from the fact that the Fock state $|N_\omega\rangle$ can be interpreted as composed of N one-photon pulses of identical frequency. Hence, all the results and considerations obtained previously apply here. An experiment that involves such a state for $N = 2$ is reported in ref. 16.

Entangled pulses of number-squeezed states combine both these

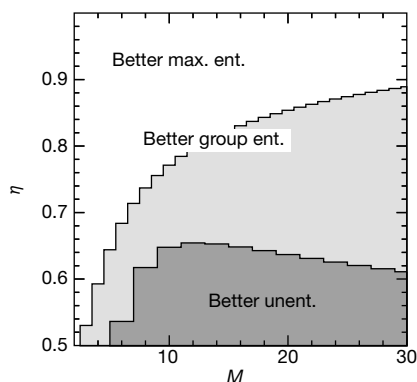


Figure 2 Sensitivity to loss. The quantum efficiency η needed to have an accuracy increase over the unentangled state $|\Psi\rangle_{un}$ is plotted versus the number M of pulses (here the number of photons per pulse is $N = 1$). The upper unshaded region is where the maximally entangled state $|\Psi\rangle_{en}$ does better than the unentangled state $|\Psi\rangle_{un}$. This unshaded region and the light grey region are where a partially entangled state, which exploits a configuration where one partially entangles subgroups of two maximally entangled photons (group entanglement), does better than $|\Psi\rangle_{un}$. The lower dark region is where the unentangled state $|\Psi\rangle_{un}$ does better.

enhancements. By replacing $|\omega\rangle$ with the number-squeezed states $|N_\omega\rangle$ in the M -fold entangled state of equation (6), one immediately obtains an improvement of $\sqrt{MN}$ over the accuracy obtainable by using M classical pulses of N photons each.

The enhanced accuracy achieved comes at the cost of an enhanced sensitivity to loss. If one or more of the photons fails to arrive, the time of arrival of the remaining photons do not convey any timing information. The simplest way to solve this problem is to ignore all trials where one or more photons is lost. A more sophisticated method is to use partially entangled states: these states provide a lower level of accuracy than fully entangled states, but are more tolerant to loss. As shown in Fig. 2, even the simple protocol of ignoring trials with loss still surpasses the unentangled-state accuracy limit even for significant loss levels. The use of intrinsically loss-tolerant, partially entangled states does even better (V.G., S.L. and L.M., unpublished results).

It is useful to consider the following intuitive picture of quantum measurements of timing. A quantum system such as a pulse of photons or a measuring apparatus with spread in energy ΔE can evolve from one state to an orthogonal state in time Δt no less than $\hbar/(4\Delta E)$ (ref. 17). Accordingly, to make more accurate timing measurements, we require states with sharp time dependence, and hence high spreads in energy. Classically, combining M systems each with spread in energy ΔE results in a joint system with spread in energy $\sqrt{M}\Delta E$. Quantum-mechanically, however, M systems can be put in entangled states in which the spread in energy is proportional to $M\Delta E$. Similarly, N photons can be joined in a squeezed state with spread in energy $N\Delta E$. The Margolus–Levitin theorem¹⁸ limits the time Δt it takes for a quantum system to evolve from one state to an orthogonal one by $\Delta t \geq \hbar/(4E)$, where E is the average energy of a system (taking the ground state energy to be 0). This result implies that the $\sqrt{MN}$ enhancement presented here is the best possible. □

Received 5 March; accepted 11 June 2001.

1. Caves, C. M. Quantum-mechanical noise in an interferometer. *Phys. Rev. D* **23**, 1693–1708 (1981).
2. Bondurant, R. S. & Shapiro, J. H. Squeezed states in phase-sensing interferometers. *Phys. Rev. D* **30**, 2548–2556 (1984).
3. Yurke, B., McCall, S. L. & Klauder, J. R. SU(2) and SU(1,1) interferometers. *Phys. Rev. A* **33**, 4033–4054 (1986).
4. Holland, M. J. & Burnett, K. Interferometric detection of optical phase shifts at the Heisenberg limit. *Phys. Rev. Lett.* **71**, 1355–1358 (1993).
5. Dowling, J. P. Correlated input-port, matter-wave interferometer: Quantum-noise limits to the atom-laser gyroscope. *Phys. Rev. A* **57**, 4736–4746 (1998).
6. Bollinger, J. J., Itano, W. M., Wineland, D. J. & Heinzen, D. J. Optimal frequency measurements with maximally correlated states. *Phys. Rev. A* **54**, R4649–R4652 (1996).
7. Boto, A. N. *et al.* Quantum interferometric optical lithography: exploiting entanglement to beat the diffraction limit. *Phys. Rev. Lett.* **85**, 2733–2736 (2000).
8. Grover, L. K. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.* **79**, 325–328 (1997).
9. Jacobson, J., Björk, G., Chuang, I. & Yamamoto, Y. Photonic de Broglie waves. *Phys. Rev. Lett.* **74**, 4853–4858 (1995).
10. Steinberg, A. M., Kwiat, P. G. & Chiao, R. Y. Dispersion cancellation in a measurement of the single-photon propagation velocity in glass. *Phys. Rev. Lett.* **68**, 2421–2424 (1992).
11. Jozsa, R., Abrams, D. S., Dowling, J. P. & Williams, C. P. Quantum clock synchronization based on shared prior entanglement. *Phys. Rev. Lett.* **85**, 2010–2013 (2000).
12. Chuang, I. L. Quantum algorithm for distributed clock synchronization. *Phys. Rev. Lett.* **85**, 2006–2009 (2000).
13. Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* (Cambridge Univ. Press, Cambridge, 1995).
14. Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* **59**, 2044–2046 (1987).
15. Hagelstein, P. L. Application of a photon configuration space model to soliton propagation in a fiber. *Phys. Rev. A* **54**, 2426–2438 (1996).
16. Fonseca, E. J. S., Monken, C. H. & Pádua, S. Measurement of the de Broglie wavelength of a multiphoton wave packet. *Phys. Rev. Lett.* **82**, 2868–2871 (1999).
17. Peres, A. *Quantum Theory: Concepts and methods* (Kluwer Academic, Dordrecht, 1993).
18. Margolus, N. & Levitin, L. B. The maximum speed of dynamical evolution. *Physica D* **120**, 188–195 (1998).

Acknowledgements

This work was funded by the ARDA, NRO, and by ARO under a MURI program.

Correspondence and requests for materials should be addressed to S.L. (e-mail: slloyd@mit.edu).

Diamagnetic activity above T_c as a precursor to superconductivity in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ thin films

Ienari Iguchi, Tetsuji Yamaguchi & Akira Sugimoto

Department of Physics and CREST-JST, Tokyo Institute of Technology,
2-12-1 Oh-okayama, Meguro-ku, Tokyo 152-8551, Japan

Superconductors show zero resistance to electric current, and expel magnetic flux (the Meissner effect) below the transition temperature (T_c). In conventional superconductors, the 'Cooper pairs' of electrons that are responsible for superconductivity form only below T_c . In the unconventional high- T_c superconductors, however, a strong electron correlation is essential for pair formation: there is evidence^{1–5} that some pairs are formed above T_c in samples that have less than the optimal density of charge carriers (underdoped) and an energy gap—the 'pseudogap'—appears to be present. Moreover, excitations that look like the vortices that carry magnetic flux inside the superconducting state have been reported above T_c (refs 6, 7). Although the origin of the pseudogap remains controversial^{8–11}, phase fluctuations above T_c , leading to some form of local superconductivity or local pairing, seem essential^{9,12–16}. Here we report magnetic imaging (scanning SQUID microscopy) of $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ thin films. Clear quantized vortex patterns are visible below T_c (18–19 K), and we observe inhomogeneous magnetic domains that persist up to 80 K. We

interpret the data as suggesting the existence of diamagnetic regions that are precursors to the Meissner state.

Scanning SQUID microscopy provides a powerful technique for observing a local, small magnetic image on the surface of a sample. The method has been used to observe the half-integer quantum vortex typical of d -wave superconductors^{17,18} and the local super-current distribution associated with the vortex state¹⁹. It should also be possible to use this technique to monitor the generation of the Meissner state in real space.

We deposited films of $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ (LSCO)—with a thickness of a few to several hundred nanometres—on $\text{SrTiO}_2(100)$ substrates by the pulsed-laser deposition method. X-ray diffraction analysis showed a strong c -axis orientation and the absence of a second phase. Energy dispersive spectroscopy analysis showed that x , the Sr component, was about 0.1, corresponding to an 'underdoped' regime. The curves of resistivity versus temperature also exhibited the behaviour expected for underdoped materials. T_c was 18–19 K. Bulk diamagnetic susceptibility was observable below 18 K using the SQUID (superconducting quantum interference device) susceptometer. The scanning SQUID microscope used here was quite similar to that reported in ref. 20. The d.c. SQUID was made of low-temperature Nb/ AlO_x /Nb tunnel junctions, and the radius of the pickup coil was 5 μm . The magnetic flux sensitivity was less than $5 \times 10^{-6} \Phi_0 \text{ Hz}^{-1/2}$, where $\Phi_0 = 2.07 \times 10^{-15} \text{ Wb}$. The separation between sample and pickup coil was 4–5 μm . A Permalloy magnetic shield was employed, yielding a background magnetic field of about 0.5 μT . A small magnetic field could be generated when the coil was wound around the sample. The sample stage was driven by three stepper motors, and the minimum step of the stage was 25 nm. The temperature of the sample stage was controllable from 3 to 90 K.

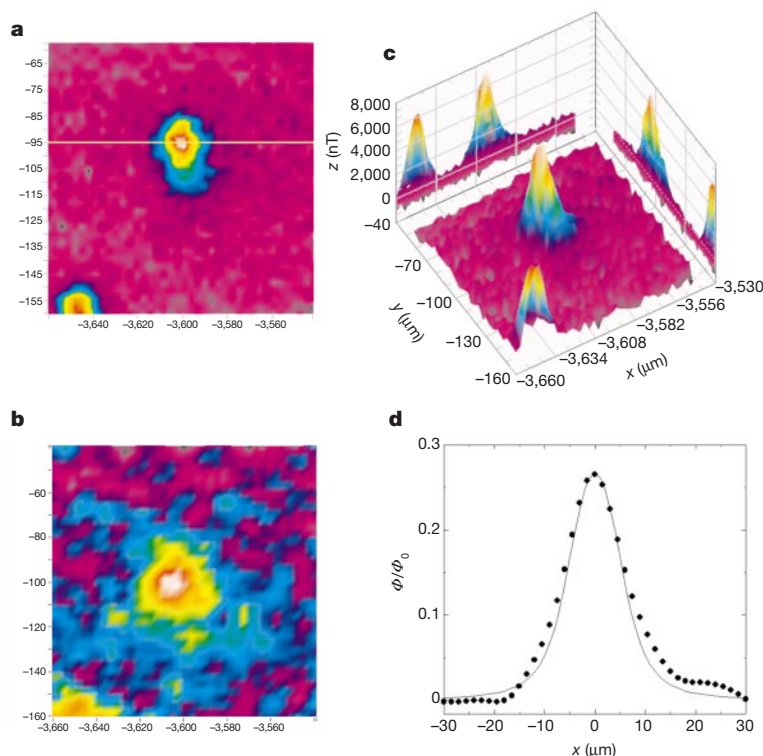


Figure 1 Scanning SQUID microscope image of quantized vortices below T_c . (In these samples, $T_c \approx 19 \text{ K}$.) **a**, **b**, Top views of the LSCO thin film at 3 K (**a**) and 14.5 K (**b**) cooled in a background magnetic field of about 0.5 μT . The pink-violet plane is the Meissner state. Two vortices are visible. The vortex image appears large (about 15 μm) in comparison with its intrinsic size ($<1 \mu\text{m}$) due to the 10- μm SQUID pickup coil and the sample-coil separation distance of about 5 μm . **c**, Three-dimensional vortex image at 3 K. **d**, Dots are the spatial magnetic-field distribution along the labelled cross-sectional

line in **a**. The solid line corresponds to the calculated field produced by single flux quantum Φ_0 , assuming a monopole source model in which the measured flux is given by $\Phi = \Phi_0/2\pi \int_C (z/r^3) dx dy$, where z is the distance between sample and pickup coil, r is the distance between the monopole source and the centre of the pickup coil, and the integration is performed over the area of the pickup coil C . The fitting was done using $z = 5.4 \mu\text{m}$. In **a** and **b**, axis units are μm .

The scanning was performed on the thin-film surface; hence, the magnetic field along the z -direction normal to the surface was dominantly detected. Because of the use of a cantilever, the pickup coil was inclined by several degrees from the horizontal plane.

Figure 1 shows the measurements on a 500-nm-thick LSCO film. At temperatures below T_c , clear vortex patterns trapped by the pinning centres were observable. Figure 1a and b show the magnetic images of the quantized vortex structure at 3 and 14.5 K, respectively. A background field of about $0.5 \mu\text{T}$ was directed upwards in the plane of the paper. Figure 1c is the three-dimensional view of the vortices observed at 3 K. We consider the basal plane (pink-violet colour) to be in the Meissner state. Figure 1d shows the observed flux in units of Φ_0 across a vortex structure. The dots are the measured data, and the solid line corresponds to the calculated curve obtained by assuming a magnetic monopole model with $z = 5.4 \mu\text{m}$. The agreement between the experiment data and the theory is reasonably good. The vortex pattern remained unchanged up to $T_c \approx 19 \text{ K}$ and disappeared above T_c . However, above T_c , the magnetic image never became uniform, and a magnetic field with inhomogeneous distribution appeared.

Figure 2 shows the top view of a magnetic image at 25.5 K. Several clear domains (pink-violet colour) with the magnetic level lower than the background field (yellow colour) were visible, and appeared to be weakly diamagnetic. The domains are nearly plateaus, as shown in the cross-sectional view. The difference in the magnetic field between the diamagnetic domain level and the background level was about $4 \mu\text{T}$ in this figure. The domain size was typically a few to several tens of micrometres at the detector level. This structure was stable at fixed temperature under low stray magnetic field, and was qualitatively similar among the samples tested.

The distribution of diamagnetic domains varied with sample temperature. Figure 3 shows a series of observed domain patterns at various temperatures for an LSCO film with a thickness of 230 nm. At higher temperatures, the nucleation of diamagnetic domains starts with the occupation of a small area. The local diamagnetic amplitude appearing here was, however, significantly larger than the

conventional Meissner screening. With reducing temperature, these domains developed to form one big domain with decreasing diamagnetic amplitude. Just after the coverage of most area by the big domain region (19 K), the superconducting transition occurred,

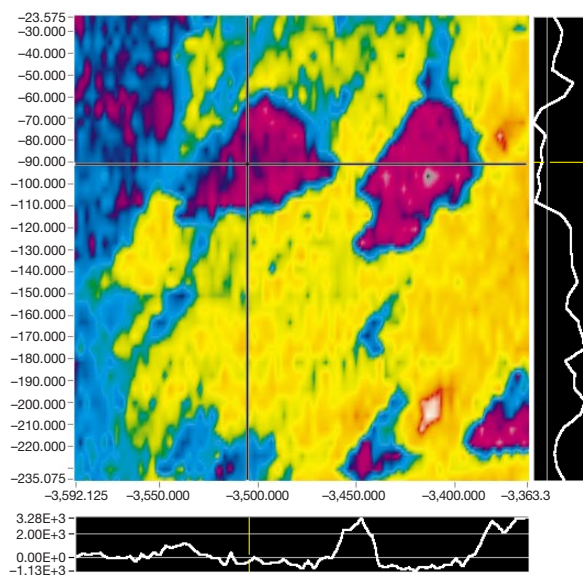


Figure 2 Scanning SQUID microscope image of magnetic domains at 25.5 K above T_c . The figure shows the diamagnetic precursor domains (pink-violet regions). The magnetic domains appeared in the opposite direction to the background field (yellow colour). As shown in the cross-sectional trace of this figure, each domain—of a few to several tens of micrometres in size—was found to be nearly plateau-shaped, quite different from the shape of the conventional vortex structure. Main figure axis units, μm ; side panel units, nT.

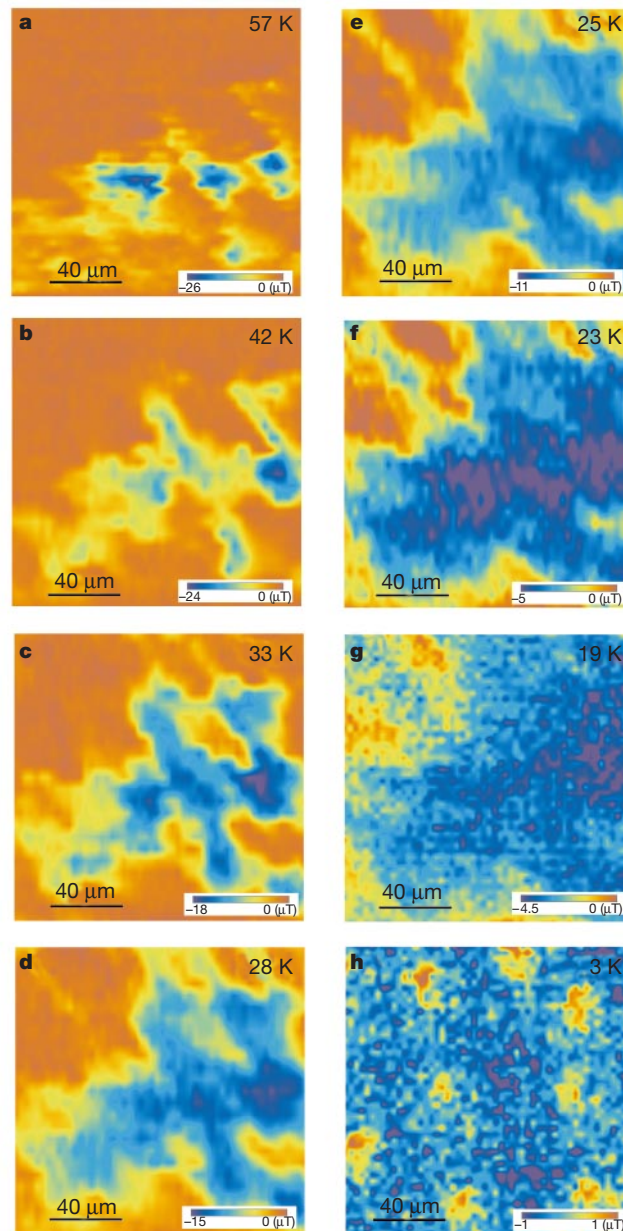


Figure 3 Development of magnetic domains with temperature (a–h), as observed by scanning SQUID microscopy. T_c was 18 K for this sample. Three or more diamagnetic precursor domains, a few tens of micrometres in size (light- and dark-blue regions, corresponding to the pink-violet region in Figs 1 and 2) are already present at 57 K. We note that the dark-orange region corresponds to the background-field level. A similar but less developed structure ($<5 \mu\text{m}$) was also observable at 80 K, which indicates that the nucleation of diamagnetic domains would start at temperatures significantly higher than 80 K. With reducing temperature, these domains developed to form one big domain, which developed further. The growth of domains became remarkable from temperatures approximately 10 K above T_c . Just after the coverage of most area by the big-domain region (19 K), the superconducting transition occurred, and the distribution of quantized vortices appeared (dark-orange spots in h). In contrast to the development of the domain area, the diamagnetic amplitude of high-temperature domains was significantly large in the first stage and then decreased with reducing temperature. We note that some local flux regions surrounded by the diamagnetic precursor domains are also observable in c–e, and the quantized vortices (orange dots) are seen in h.

and the distribution of quantized vortices appeared. In Fig. 4, we show as a function of temperature the maximum diamagnetic peak height (Fig. 4a), the fractional area of domains (Fig. 4b), and the total diamagnetic flux involved in the fractional area (Fig. 4c). The peak diamagnetic amplitude became largest around 55 K, which corresponds to several times the flux quantum Φ_0 . The peak diamagnetic amplitude rapidly decreased towards the Meissner level as the temperature approached T_c . We note that the zero level corresponds to the Meissner state ($B = 0$). The signal seemed to decrease again above 60 K, with an almost vanishing fractional area. Hence, we consider the magnetic domain—that is, the ‘diamagnetic domain’ that appeared above T_c —to be a diamagnetic precursor state to the Meissner state, produced by performed pairs of electrons. We also observed some local positive magnetic flux regions surrounded by the precursor Meissner domain (Fig. 3c–e).

We also performed measurements on a nearly optimally doped LSCO single crystal (thickness 0.5 mm, $x = 0.135$) with a carefully polished surface. Whereas clear quantized vortices were imaged below T_c , similar diamagnetic precursor domains were also observable at 45 K—that is, 5 K above the T_c of ~ 40 K. For low- T_c niobium thin films, however, no such signal was obtained above T_c . On the other hand, for a slightly underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ thin film ($T_c = 84$ K), such domains were again observable in the limited,

narrow temperature range. These results will be published elsewhere.

Because the measurements by scanning SQUID microscope showed continuous transition to the Meissner state, the diamagnetic precursor state to the Meissner state really exists at temperatures far above T_c . Initially, some nucleation of the precursor state with large diamagnetic amplitude seems to occur locally. It is not clear at present whether this state appears intrinsically or is caused by some crystal defects, although it seems to appear randomly. Even if this state is due to inhomogeneity or impurities, it is essential in that it leads to the high- T_c superconductivity. The observed precursor state may not be simply interpreted as a spin-singlet state such as that given by the t - J model in which the energy balance of the system is described by the transfer integral of electrons (t) and the superexchange interaction of Cu spins (J), as it is spatially distributed. The appearance of a spatially varying state is consistent with recent experimental observations based on the Nernst effect^{6,7}. If there is a temperature gradient in the film, the spatial pattern will change through the film, leading to the possible occurrence of a certain kind of local flux flow. However, two experiments measured qualitatively different quantities. Xu *et al.*⁶ measured the dynamic flux flow state, which provides indirect information on the spatially varying state. On the other hand, our experiment measured directly the spatial distribution of a local stationary diamagnetic state. In the conventional superconducting state, the diamagnetic behaviour is directly connected with the screening current, hence quantized vortices. However, above T_c , the Cooper pairs will not be in the coherent state. So the connection between the diamagnetic property and vortex movement is unclear.

Our results are consistent with the existence of a pseudogap state, as the presence of localized domains may somehow contribute to the creation of such a state. The pseudogap may correspond to the precursor gap state before the appearance of a well-developed superconducting gap state. If the observed diamagnetic precursor domains are associated with the pseudogap state, it suggests a random nucleation of the pseudogap regions in space, which may be attributed to the extremely short coherence length in high- T_c superconductors. Stripe structures^{8,21,22} seem not to play an important role in the formation of observed diamagnetic precursor state as the length and time scales of measurements are much longer than would be expected from stripe formation. On the contrary, our results suggest the existence of an intrinsic tendency to phase separation, such as reported for doped antiferromagnets²³. The possible occurrence of an intrinsic tendency toward electronic inhomogeneity on multiple length scales has been discussed in ref. 24. $\square$

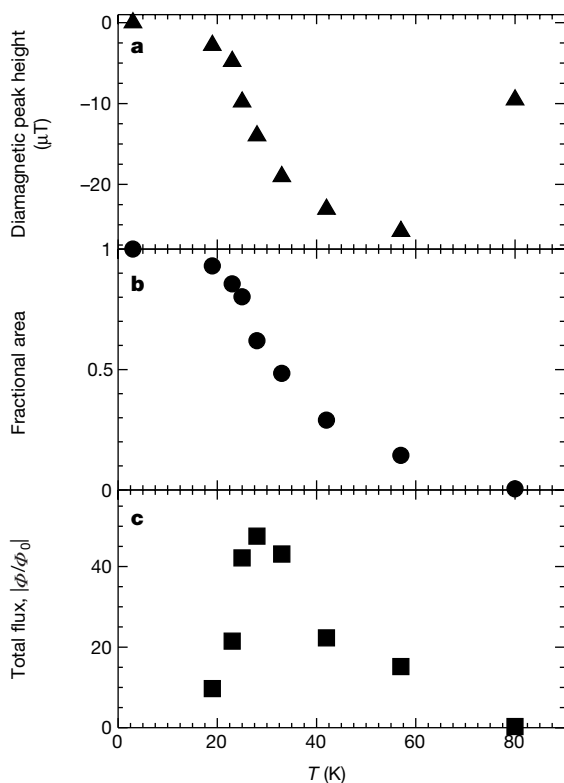


Figure 4 Development of quantitative scales in the magnetic domains with temperature. **a**, The maximum diamagnetic amplitude versus temperature; **b**, the fractional area of the magnetic domains versus temperature; and **c**, the total integrated diamagnetic flux in the fractional area versus temperature. With reducing temperature, first, diamagnetic nucleation occurs above 80 K and tiny domains appear. With further reduction of temperature, the diamagnetic amplitude increases, then decreases toward T_c , forming a downward peak structure with its maximum peak height around 55 K. On the other hand, the fractional area of domains increases monotonically toward T_c . The total integrated flux inside the fractional area exhibits complicated behaviour, with its peak just below 30 K. However, the total integrated flux simply provides a measure of the total flux inside the area of the scanning window. We note that the data point at 80 K was recorded in a slightly displaced area.

Received 11 December 2000; accepted 25 May 2001.

- Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62**, 61–122 (1999).
- Loeser, A. G. *et al.* Excitation gap in the normal state of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Science* **273**, 325–329 (1996).
- Ding, H. *et al.* Spectroscopic evidence for a pseudogap in the normal state of underdoped high- T_c superconductors. *Nature* **382**, 51–54 (1996).
- Imai, T. *et al.* Cu spin dynamics in high T_c and related oxides investigated by nuclear spin-lattice relaxation. *Physica C* **162–164**, 169–170 (1989).
- Corson, J., Mallozzi, R., Orenstein, J., Eckstein, J. N. & Bozovic, I. Vanishing of phase coherence in underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Nature* **398**, 221–223 (1999).
- Xu, Z. A., Ong, N. P., Wang, Y., Kakeshita, T. & Uchida, S. Vortex-like excitations and the onset of superconducting phase fluctuation in underdoped $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Nature* **406**, 486–488 (2000).
- Lee, P. A. Some vortices like it hot. *Nature* **406**, 467–468 (2000).
- Orenstein, J. & Millis, A. J. Advances in the physics of high-temperature superconductivity. *Science* **288**, 468–474 (2000).
- Emery, V. J. & Kivelson, S. A. Importance of phase fluctuations in superconductors with small superfluid density. *Nature* **374**, 434–437 (1995).
- Moriya, T., Takahashi, Y. & Ueda, K. Antiferromagnetic spin fluctuations and superconductivity in two-dimensional metals—a possible model for high T_c oxides. *J. Phys. Soc. Jpn* **59**, 2905–2915 (1990).
- Monthoux, P., Balatsky, A. V. & Pines, D. Toward a theory of high-temperature superconductivity in the antiferromagnetically correlated cuprate oxides. *Phys. Rev. Lett.* **67**, 3448–3451 (1991).
- Doniach, S. & Inui, M. Long-range Coulomb interactions and the onset of superconductivity in the high- T_c materials. *Phys. Rev. B* **41**, 6668–6678 (1990).

13. Roddick, E. & Stroud, D. Effect of phase fluctuations on the low-temperature penetration depth of high- T_c superconductors. *Phys. Rev. Lett.* **74**, 1430–1433 (1995).
14. Roshnar, D. & Kivelson, S. A. Superconductivity and the quantum hard-core dimer gas. *Phys. Rev. Lett.* **61**, 2376–2379 (1988).
15. Sá de Melo, C. A. R., Randeria, M. & Engelbrecht, J. R. Crossover from BCS to Bose superconductivity: transition temperature and time-dependent Ginzburg–Landau theory. *Phys. Rev. Lett.* **71**, 3202–3205 (1993).
16. Carlson, E. W., Kivelson, S. A., Emery, V. J. & Manousakis, E. Classical phase fluctuations in high temperature superconductors. *Phys. Rev. Lett.* **83**, 612–615 (1999).
17. Tsuei, C. C. *et al.* Pairing symmetry and flux quantization in tricrystal superconducting ring of $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Phys. Rev. Lett.* **73**, 592–596 (1994).
18. Tsuei, C. C. & Kirtley, J. R. Pairing symmetry in cuprate superconductors. *Rev. Mod. Phys.* **72**, 969–1016 (2000).
19. Sugimoto, A., Yamaguchi, T. & Iguchi, I. Supercurrent distribution in high- T_c superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ thin films by scanning superconducting quantum interference device microscopy. *Appl. Phys. Lett.* **77**, 3069–3071 (2000).
20. Morooka, T., Nakayama, S., Odawara, A. & Chinone, K. Observation of superconducting device using magnetic imaging system with a micro-DC superconducting quantum interference device magnetometer. *Jpn J. Appl. Phys.* **38**, L119–L122 (1999).
21. Emery, V. J., Kivelson, S. A. & Tranquada, J. M. Stripe phases in high-temperature superconductors. *Proc. Natl Acad. Sci. USA* **96**, 8814–8817 (1999).
22. Zaanen, J. Self-organized one dimensionality. *Science* **286**, 251–252 (1999).
23. Kiyukhin, V. *et al.* Multiphase segregation and metal–insulator transition in single crystal $\text{La}_{0.8-x}\text{Pr}_x\text{Ca}_{0.2}\text{MnO}_3$. *Phys. Rev. B* **63**, 24420–1–24420–7 (2000).
24. Emery, V. J. & Kivelson, S. A. Frustrated electronic phase separation and high-temperature superconductors. *Physica C* **209**, 597–621 (1993).

Acknowledgements

We thank M. Ueda and M. Koyanagi for discussions, and I. Tanaka for providing LSCO single crystals. This work was supported by the CREST program (Core Research for Evolutional Science and Technology) of the Japan Science and Technology Corporation (JST).

Correspondence and requests for materials should be addressed to I.I. (e-mail: iguchi@ap.titech.ac.jp).

Sn-zeolite beta as a heterogeneous chemoselective catalyst for Baeyer–Villiger oxidations

Avelino Corma*, Laszlo T. Nemeth†, Michael Renz* & Susana Valencia*

* Instituto de Tecnología Química, UPV-CSIC, Avda. de los Naranjos s/n, 46022 Valencia, Spain

† Universal Oil Products, UOP LLC, 50 East Algonquin Road, Des Plaines, Illinois 60017-5016, USA

The Baeyer–Villiger oxidation, first reported more than 100 years ago¹, has evolved into a versatile reaction widely used² to convert ketones—readily available building blocks in organic chemistry—into more complex and valuable esters and lactones. Catalytic versions of the Baeyer–Villiger oxidation are particularly attractive for practical applications, because catalytic transformations simplify processing conditions while minimizing reactant use as well as waste production. Further benefits are expected from replacing peracids, the traditionally used oxidant, by cheaper and less polluting hydrogen peroxide³. Dissolved platinum complexes⁴ and solid acids, such as zeolites^{5,6} or sulphonated resins⁷, efficiently activate ketone oxidation by hydrogen peroxide. But these catalysts lack sufficient selectivity for the desired product if the starting material contains functional groups other than the ketone group; they perform especially poorly in the presence of carbon–carbon double bonds. Here we show that upon incorporation of 1.6 weight per cent tin into its framework, zeolite beta acts as an efficient and stable heterogeneous catalyst for the Baeyer–Villiger oxidation of saturated as well as unsaturated ketones by hydrogen peroxide, with the desired lactones

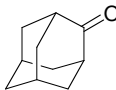
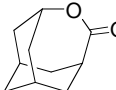
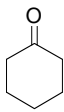
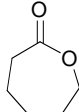
forming more than 98% of the reaction products. We ascribe this high selectivity to direct activation of the ketone group, whereas other catalysts first activate hydrogen peroxide, which can then interact with the ketone group as well as other functional groups.

When trying to avoid the use of peracids for the Baeyer–Villiger reaction, the methodology developed up to now involved catalysts able to activate hydrogen peroxide, H_2O_2 . Amongst homogeneous catalysts, complexes of molybdenum⁸ and rhenium⁹ have been shown to activate H_2O_2 , but their turnover numbers (TONs) and selectivities are relatively low (TONs are below 20 for overall reaction times ranging from 12 to 24 hours). Pt complexes achieve TONs of about 50 within ~5 hours, but are not chemoselective when other functional groups are present¹⁰. Concerning heterogeneous catalysts, supported Pt complexes¹¹ and TS-1¹² have been used, but have shown only limited activity and selectivity, respectively. Acid zeolites such as H-beta and USY activate hydrogen peroxide for the Baeyer–Villiger oxidation, but show selectivities of less than 60–70% (ref. 5). MeReO_3 (ref. 13), TS-1³ and Pt complexes¹⁴ are excellent epoxidation catalysts in the presence of H_2O_2 and consequently, epoxidation is favoured over the Baeyer–Villiger oxidation when using unsaturated ketones as starting material.

Attempts to develop efficient Baeyer–Villiger oxidation catalysts have so far focused on catalysts that increase the nucleophilicity of H_2O_2 in order to facilitate attack by the oxidizing species on the carbonyl carbon atom⁴. We decided to pursue a different concept that involves initial selective activation of the carbonyl group, followed by reaction with non-activated H_2O_2 . Lewis acids^{14,15} are known to activate carbonyl groups, and we therefore focused on designing a solid catalyst with isolated and uniform Lewis-acid sites, which could activate the carbonyl group in preference to H_2O_2 . We chose zeolite beta as the solid catalyst matrix because it allows the incorporation of transition metals in framework positions^{16,17}, and Sn as the Lewis acid. (The list of potential Lewis acids considered included Sn, Ti, V, Cr and Fe. Sn was selected because Ti, V and Cr are well known catalysts for activating H_2O_2 and performing epoxidation of olefins, while Fe can decompose H_2O_2 into H_2O and oxygen¹⁸).

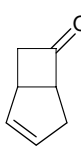
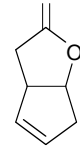
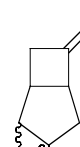
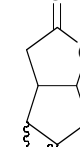
We synthesized Sn-zeolite beta, which provides a catalyst with discrete and isolated Lewis sites, in fluoride medium¹⁹. Tetraethyl-orthosilicate (TEOS) was hydrolysed in an aqueous solution of tetraethylammonium hydroxide (TEAOH) under stirring. Then a solution of $\text{SnCl}_4 \cdot 5 \text{H}_2\text{O}$ in water was added, and the mixture was stirred until the ethanol formed upon hydrolysis of TEOS was evaporated. HF was added to the clear solution obtained, and a thick paste was formed. Finally, an aqueous suspension of ‘seeds’ of dealuminated nanocrystalline (20 nm) zeolite beta was also added.

Table 1 Baeyer–Villiger oxidation of adamantanone and cyclohexanone catalysed by Sn-zeolite beta

Substrate	Product	Reaction conditions	Yield (%)	Selectivity (%)
		MTBE, 56 °C, 6 h	94	>98
		Dioxane, 90 °C, 3 h	52	>98

H_2O_2 (35 wt% H_2O_2 in H_2O) was used as oxidant (ketone/ H_2O_2 was 1.0 mol mol⁻¹ for adamantanone (top row), and 1.5 mol mol⁻¹ for cyclohexanone (bottom row). Methyl *t*-butyl ether (MTBE) or dioxane were used as solvent (as shown solvent/ketone was 20 g g⁻¹ and 30 g g⁻¹ for adamantanone and cyclohexanone, respectively). 1.5 wt% of Sn-zeolite beta catalyst was employed with respect to the total solution, which was equivalent to 0.66 mol% of Sn with respect to the ketone.

Table 2 Baeyer–Villiger oxidation of bicyclo[3.2.0]hept-2-en-6-one using different oxidation systems

	Reactant conversion (%)	Product selectivity (%)			
					
Sn-zeolite	>95	100*	0	0	0
mCPBA†	>95	29	34	37	0
Enzymes ²⁰		100*	0	0	0

The reaction was performed employing the amounts given in Table 1, in MTBE as solvent at 30 °C. Reaction time, 2 h.

* As 67 : 33 mixture of isomers.

† *m*-chloroperbenzoic acid.

The final gel composition was as follows: 1.0 SiO₂ : 0.0083 SnO₂ : 0.54 TEAOH : 7.5 H₂O : 0.54 HF.

We performed the crystallization in Teflon-lined stainless steel autoclaves, which were heated to 140 °C and rotated over a period of 20 days. X-ray diffraction of the dried (100 °C overnight) and calcined (580 °C for 3 hours) product showed it to be a highly crystalline zeolite beta (see Supplementary Information), which—according to chemical analysis—contained 1.6 wt% Sn. ¹¹⁹Sn magic-angle-spinning NMR of a Sn-zeolite beta enriched with the ¹¹⁹Sn isotope confirmed that Sn is incorporated into the zeolite framework as tetrahedrally coordinated Sn (−444 p.p.m.; see Supplementary Information).

In situ measurements using infrared spectroscopy (see Supplementary Information) indicated that Sn-zeolite beta was indeed able to activate cyclohexanone at 25, 50, 100 and 200 °C, reflected by a shift of 48 cm^{−1} in the signal of the carbonyl group towards lower wavenumbers. When using Ti-zeolite beta and V-zeolite beta, much smaller shifts (~25 cm^{−1}) occur at 100 °C; when heating at and beyond 200 °C, the lack of bands in the signal indicates the lack of interaction, revealing that cyclohexanone interacts far less strongly with Ti-zeolite beta and V-zeolite beta than with Sn-zeolite beta.

We used Sn-zeolite beta to catalyse the Baeyer–Villiger oxidations of adamantanone and cyclohexanone in a batch reactor, under stirring at 56 and 90 °C, respectively. The results, summarized in Table 1, show excellent selectivity for lactone formation; the conversions correspond to TONs of 116 over 6 hours reaction time, and

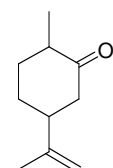
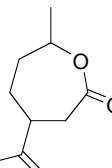
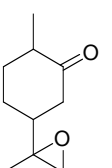
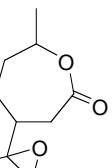
88 over 3 hours of reaction time, for the transformation of adamantanone and cyclohexanone, respectively. After completion of reaction in one batch, the catalyst was filtered out and reused; its activity remained at a constant level, giving after four reaction cycles (each again lasting 6 and 3 hours, respectively) TONs of 430 and 345 for adamantanone and cyclohexanone, respectively. These values are almost an order of magnitude higher than the best homogeneous catalytic system¹⁰.

We confirmed that reaction does not occur in the absence of the catalyst, and that it is not the result of homogeneous catalysis by leached Sn. (After reaching a conversion of up to 30%, the solid catalyst was filtered out at the reaction temperature. This stopped the reaction, with no further conversion being observed within the next four hours.) We also used SnCl₄ as catalyst for cyclohexanone oxidation and obtained a TON of 5; the analogous result with Sn-zeolite beta is a TON of 88, under identical reaction conditions. Neither SiO₂ (Aerosil, 200 m² g^{−1}), nor pure silica zeolite beta impregnated with a SnCl₄ solution, were catalytically active for the above reaction. These results emphasize that efficient and selective catalysis requires Sn to be very highly dispersed, and most likely also that it needs to be situated within the zeolite framework. Dimeric peroxides, common by-products for the cyclohexanone oxidation by hydrogen peroxide¹⁴, were not detected (by thin layer chromatography, detection limit <1%) when comparing with an authentic sample of 7,8,15,16-tetraoxa-dispiro[5,2,5,2]hexadecane.

Our Sn-zeolite beta catalyst retains its high chemoselectivity when carrying out the Baeyer–Villiger oxidation of unsaturated ketones. As illustrated by the data in Table 2, the Baeyer–Villiger oxidation of bicyclohept-3-en-1-one using *m*-chloroperbenzoic acid (mCPBA) as oxidant, which is common practice, leads to the production of lactones and epoxide by-products, whereas Sn-zeolite beta yields only two regioisomeric lactones. This 100% selectivity for lactones has previously only been obtained when using enzymes as catalysts²⁰.

The oxidation of dihydrocarvone further illustrates the chemoselectivity achieved with our catalyst (see Table 3). With Sn-zeolite

Table 3 Baeyer–Villiger oxidation of dihydrocarvone using different oxidation systems

Oxidant	Reactant conversion (%)	Product selectivity (%)			
					
Sn-zeolite beta/H ₂ O ₂	68	100	0	0	0
mCPBA†	85	11	71	18	0
Ti-zeolite beta/H ₂ O ₂	48	0	79*	0	0

The same conditions as for adamantanone (see Table 1) were employed.

* Rest missing to 100% were opening products of the epoxide.

† *m*-chloroperbenzoic acid.

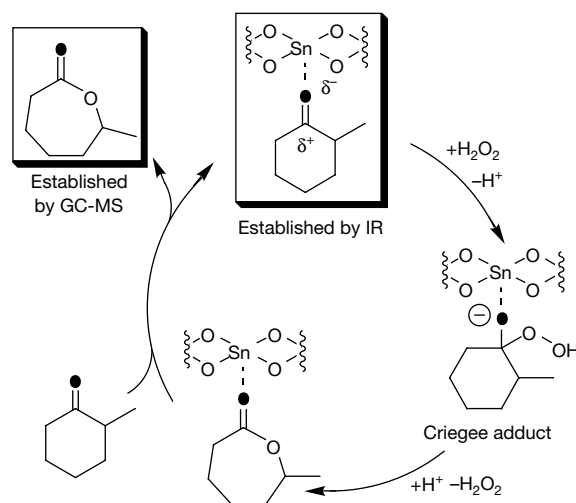


Figure 1 Catalytic cycle for the Baeyer–Villiger oxidation, using hydrogen peroxide catalysed by Sn-zeolite beta. First, the ketone is coordinated to the Lewis-acid centre and, thereby, the carbonyl group is activated. (We have shown this by *in situ* infrared spectroscopy.) Then the hydrogen peroxide attacks the more electrophilic carbonyl carbon atom. After the rearrangement step, the lactone product is replaced by a new substrate molecule. This mechanism is analogous to that for peracids, and we have confirmed it by means of an ¹⁸O labelling experiment (see text). IR, infrared spectroscopy; GC-MS, gas chromatography-mass spectrometry.

beta catalyst, only the corresponding lactone is produced, while Ti-zeolite beta catalyst produced epoxide product and the corresponding diols, and the peracid (mCPBA) yielded a mixture of lactone, epoxide and epoxylactone.

Our infrared observations (see Supplementary Information) imply that the first step of the reaction involves the activation of the carbonyl group of the ketone on the Sn atoms of the Sn-zeolite beta. To elucidate the reaction mechanism further, we used 2-methylcyclohexanone labelled with ^{18}O (^{18}O content 89%) as reactant; monitoring the products with gas chromatography–mass spectrometry established that the Baeyer–Villiger oxidation leads to incorporation of all labelled oxygen into the lactone product as carbonyl oxygen, and that none is present as ring oxygen (see Supplementary Information). The Baeyer–Villiger oxidation in our catalytic system therefore proceeds via a ‘Criegee’ adduct, where H_2O_2 adds to the ketone activated by the Sn-zeolite beta, and the formation of dioxiranes or carbonyl oxides as intermediates can be excluded (compare ref. 2). The complete mechanistic cycle is depicted in Fig. 1. Although coordination of the hydrogen peroxide to the Sn centre and subsequent activation by deprotonation cannot be excluded, diffuse-reflectance–ultraviolet–visible spectroscopy failed to detect the Sn–O–O–H species that may be expected in such a mechanism. □

Received 13 December 2000; accepted 31 May 2001.

1. Baeyer, A. & Villiger, V. Einwirkung des Caro'schen Reagens auf Ketone. *Chem. Ber.* **32**, 3625–3633 (1899).
2. Renz, M. & Meunier, B. 100 years of Baeyer–Villiger oxidations. *Eur. J. Org. Chem.* 737–750 (1999).
3. Arends, I. W. C. E., Sheldon, R. A., Wallau, M. & Schuchardt, U. Oxidative transformations of organic compounds mediated by redox molecular sieves. *Angew. Chem. Int. Edn Engl.* **36**, 1145–1163 (1997).
4. Strukul, G. Transition metal catalysis in the Baeyer–Villiger oxidation of ketones. *Angew. Chem. Int. Edn Engl.* **37**, 1198–1209 (1998).
5. Fischer, J. & Hölderich, W. F. Baeyer–Villiger-oxidation of cyclopentanone with aqueous hydrogen peroxide by acid heterogeneous catalysis. *Appl. Catal. A* **180**, 435–443 (1999).
6. Chang, C. D. & Helling, S. D. Production of lactones and omega-hydroxycarboxylic acids. US Patent No. 4870192 (1996).
7. Hoelderich, W., Fischer, J., Schindler, G. -P. & Arntz, D. Preparation of lactones by Baeyer–Villiger oxidation of cyclic ketones. German Patent DE 19745442 (1999).
8. Jacobson, S. E., Tang, R. & Mares, F. Oxidation of cyclic ketones by hydrogen peroxide catalysed by group 6 metal peroxo complexes. *J. Chem. Soc. Chem. Commun.* 888–889 (1978).
9. Herrmann, W. A., Fischer, R. W. & Correia, J. D. G. Multiple bonds between main-group elements and transition metals. Part 133. Methyltrioxorhenium as a catalyst of the Baeyer–Villiger oxidation. *J. Mol. Catal.* **94**, 213–223 (1994).
10. Gavagnin, R., Cataldo, M., Pinna, F. & Strukul, G. Diphosphine–palladium and -platinum complexes as catalysts for the Baeyer–Villiger oxidation of ketones: effect of the diphosphine, oxidation of acyclic ketones, and mechanistic studies. *Organometallics* **17**, 661–667 (1998).
11. Palazzi, C., Pinna, F. & Strukul, G. Polymer-anchored platinum complexes as catalysts for the Baeyer–Villiger oxidation of ketones: preparation and catalytic properties. *J. Mol. Catal. A* **151**, 245–252 (2000).
12. Bhaumik, A., Kumar, P. & Kumar, R. Baeyer–Villiger rearrangement catalysed by titanium silicate molecular sieve (TS-1)/ H_2O_2 system. *Catal. Lett.* **40**, 47–50 (1996).
13. Herrmann, W. A. Essays on organometallic chemistry. VII. Laboratory curiosities of yesterday, catalysts of tomorrow: organometallic oxides. *J. Organomet. Chem.* **500**, 149–150 (1995).
14. Frisone, M. D. T., Pinna, F. & Strukul, G. Baeyer–Villiger oxidation of cyclic ketones with hydrogen peroxide catalyzed by cationic complexes of platinum(II): selectivity properties and mechanistic studies. *Organometallics* **12**, 148–156 (1993).
15. Shambayati, S. & Schreiber, S. L. in *Comprehensive Organic Synthesis* Vol. 1 (eds Trost, B. M. & Fleming, I.) 283–324 (Pergamon, Oxford, 1991).
16. Cambor, M. A., Corma, A., Martínez, A. & Pérez-Pariante, J. Synthesis of a titaniumsilicoaluminate isomorphous to zeolite beta and its application as a catalyst for the selective oxidation of large organic molecules. *J. Chem. Soc. Chem. Commun.* 589–590 (1992).
17. Sen, T., Chatterjee, M. & Sivasanker, S. Novel large-pore vanadium aluminosilicate and borosilicates with BEA structure. *J. Chem. Soc. Chem. Commun.* 207–208 (1995).
18. Huang, H. -H., Lu, M. -C. & Chen, J. -N. Catalytic decomposition of hydrogen peroxide and 2-chlorophenol with iron oxides. *Wat. Res.* **35**, 2291–2299 (2001).
19. Valencia, S. & Corma, A. Stannosilicate molecular sieves. US Patent No. 5968473A (1999).
20. Kelly, D. R., Wan, P. W. H. & Tang, J. *Biotechnology* Vol. 8a Ch. 11, 535–587 (Wiley-VCH, Weinheim, 1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank CICYT and UOP for financial support.

Correspondence and requests for materials should be addressed to A.C. (e-mail: acorma@itq.upv.es).

Warm tropical ocean surface and global anoxia during the mid-Cretaceous period

Paul A. Wilson* & Richard D. Norris†

* Southampton Oceanography Centre, School of Ocean & Earth Sciences, European Way, Southampton SO14 3ZH, UK

† Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543-1541, USA

The middle of the Cretaceous period (about 120 to 80 Myr ago) was a time of unusually warm polar temperatures¹, repeated reef-drowning in the tropics² and a series of oceanic anoxic events (OAEs) that promoted both the widespread deposition of organic-carbon-rich marine sediments and high biological turnover^{3–8}. The cause of the warm temperatures is unproven but widely attributed to high levels of atmospheric greenhouse gases such as carbon dioxide^{7–12}. In contrast, there is no consensus on the climatic causes and effects of the OAEs, with both high biological productivity and ocean ‘stagnation’ being invoked as the cause of ocean anoxia^{3–8}. Here we show, using stable isotope records from multiple species of well-preserved foraminifera, that the thermal structure of surface waters in the western tropical Atlantic Ocean underwent pronounced variability about 100 Myr ago, with maximum sea surface temperatures 3–5 °C warmer than today. This variability culminated in a collapse of upper-ocean stratification during OAE-1d (the ‘Brestroffer’ event), a globally significant period of organic-carbon burial that we show to have fundamental, stratigraphically valuable, geochemical similarities to the main OAEs of the Mesozoic era. Our records are consistent with greenhouse forcing being responsible for the warm temperatures, but are inconsistent both with explanations for OAEs based on ocean stagnation, and with the traditional view (reviewed in ref. 12) that past warm periods were more stable than today's climate.

Diverse sets of biotic and marine geological data provide convincing evidence that the mid-Cretaceous oceans were both extremely warm by today's standards and prone to repeated widespread OAEs (Fig. 1). Debate surrounding black-shale formation during OAEs has traditionally been divided along two main lines⁶. (1) ‘High productivity’ models invoke rapid supply of organic carbon to the sea floor, initiated by strong upwelling or monsoon overturning, to promote porewater anoxia. (2) ‘Ocean stagnation’ models invoke slow renewal of deep water to promote anoxia. Suggested mechanisms for stagnation include the development of intense vertical gradients of temperature and salinity, as in Pliocene–Pleistocene Mediterranean sapropel formation, and are rooted in the perception that mid-Cretaceous warmth must have been accompanied by slow ocean circulation and limited ventilation of intermediate and deep waters. In turn, extreme mid-Cretaceous warmth is commonly ascribed to high partial pressures of atmospheric greenhouse gases^{7–12}. Yet, our persistent inability to reconstruct the basic properties of Cretaceous sea water has severely limited our capacity to test these hypotheses.

In particular, we need to improve existing $\delta^{18}\text{O}$ -derived palaeotemperature records, which are based largely on measurements of bulk carbonates and/or sampling at very low temporal resolution. A handful of studies provide reliable $\delta^{18}\text{O}$ -derived temperatures in support of extreme mid-Cretaceous high-latitude warmth (for example, sea surface temperatures, SSTs, of about 15 °C in the Late Albian Antarctic)¹³, but with few exceptions, $\delta^{18}\text{O}$ records suggest that contemporaneous low-latitude SSTs were generally no warmer or significantly cooler (by up to ~12 °C) than today^{11,12,14}. Such cool tropical SSTs are problematic, because general circulation

model (GCM) experiments suggest that the high values of carbon dioxide partial pressure (2–4 times the present values) needed to explain mid-Cretaceous polar warmth should also have promoted low-latitude warming (tropical SSTs about 2–5 °C warmer than today)^{9–12}.

Ocean Drilling Program (ODP) Site 1052 provides an expanded record (sedimentation rates, ~9–10 cm kyr⁻¹) through the core of the inferred mid-Cretaceous greenhouse climate (late Albian to early Cenomanian, approximately 100 Myr ago). This site is located in the western North Atlantic (palaeolatitude, ~23 °N; Fig. 1) and consists of approximately 210 m of predominantly carbonaceous mudstones with a cyclical series of laminated black shales and clay-rich carbonates, with occasional thin limestone beds (Fig. 2). The shale–carbonate cycles correspond to pronounced cycles in sediment colour, neutron porosity and resistivity. Sediments from Site 1052 contain exceptionally well preserved foraminifera (having transparent shells with empty chambers and detailed microstructure) that co-exist with ammonite shells that retain primary nacreous metastable aragonite¹⁵. We attribute such remarkable preservation to the clay-rich lithologies sampled. We have taken care to avoid analysis of less well preserved material from the few thin limestone ‘stringers’, but nevertheless have sampled across enough carbonate–shale couplets to ensure that our results are not strongly biased towards the different environments represented by these lithologies.

The black-shale cycles in ODP Site 1052 are remarkably similar in appearance and biostratigraphically correlative to those in the Vocontion basin (eastern North Atlantic) that have been used, in conjunction with increased turnover rates in planktic foraminifera and radiolaria, to infer¹⁶ the existence of a latest Albian Breistroffer event or OAE-1d. Our site-by-site compilation confirms this view (Fig. 1). However, we need to know whether OAE-1d mirrors its precursor early Albian OAE-1b (recently interpreted as a regional ‘mega-sapropel’¹⁷) or, alternatively, represents a major perturbation to the global carbon cycle like the giant Mesozoic Cenomanian/Turonian, Aptian and Toarcian OAEs^{5,18–20}.

We have developed a chronology for the mid-Cretaceous sequence from Site 1052 using an orbital cyclostratigraphy supported by foraminifer and calcareous nannofossil biostratigraphy. The neutron porosity down-hole log delineates a regular 1.8–2.3-m cycle of clay-rich carbonate (or limestone) and shale superimposed upon a 9–10-m cycle in shales and thin clayey carbonate (or limestone) couplets (Fig. 2). A distinctive 5:1 bundling of limestone–shale cycles near the Albian/Cenomanian boundary is reminiscent of the 21-kyr orbital precession cycle bundled into ~100-kyr eccentricity cycles (Fig. 2). Assuming that the ~9–10-m cycle corresponds to the short (100-kyr) orbital eccentricity cycle, we estimate that the interval between the first appearances of the foraminifer *Rotalipora appenninica* and the calcareous nannofossil *Eiffellithus turrisseiffelii* in Site 1052 represents ~1.3 Myr, comparing favourably with the biostratigraphic estimate (1.4 Myr)²¹.

To reconstruct surface hydrography, we analysed narrow size-fractions (212–300 µm) of planktic foraminifera for which we have already determined¹⁵ depth habitats. Records of δ¹³C for surface and deeper dwelling species are notably consistent (Fig. 2). This result may indicate that these foraminifera inhabited waters with a weak vertical δ¹³C gradient or reflect the absence of photosymbiotic species¹⁵. The coherency of the signal among species (Fig. 2) suggests that change in ocean chemistry, rather than foraminifer habitat, must dominate the record. Accompanying δ¹⁸O records give a clear demonstration of depth habitat stratification among species. Surface dwellers show pronounced variability (~0.5–1.25‰ versus PDB) compared to their deeper-living counterparts (Fig. 2). This variability must reflect changes in primary δ¹⁸O values (rather than differential diagenetic alteration) given the exceptional preservation of the foraminifera. The large magnitude of the variations in the surface record compared to the ‘thermocline’ record (Fig. 2) and the open ocean setting (Fig. 1) suggest that most of the signal is caused by changes in surface temperature (rather than seawater δ¹⁸O) at the site. Assuming reasonable, conservative estimates for seawater δ¹⁸O (see Fig. 2 legend) and sources of error²², we estimate that Albian

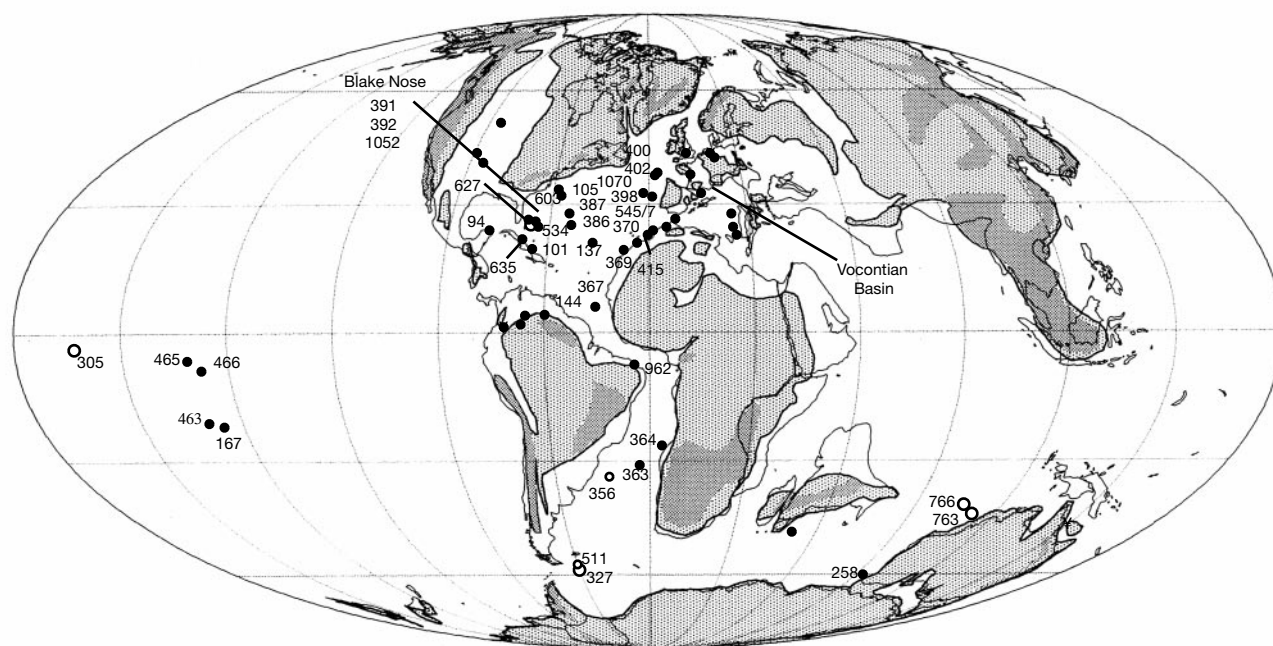


Figure 1 Reconstruction of the mid-Cretaceous (Albian)²⁷ period, showing the global location of sites with well-dated sections correlative to the upper Albian Breistroffer oceanic anoxic event. This event, OAE-1d, is defined by the *Rotalipora appenninica* planktonic foraminifera zone. Solid circles, sites that show some lithological expression of

OAE-1d (from subtle lamination through dark carbonaceous mudstones to well-developed black shales, total organic-carbon contents up to ~25 wt%; ref. 28). Open circles, sites that show no obvious lithological expression of OAE-1d. Numbers refer to Deep Sea Drilling Project and Ocean Drilling Program Sites.

SSTs in the western tropical Atlantic underwent significant variation, with maximum changes of at least 6 °C on ~10-kyr timescales. To our knowledge, the minimum $\delta^{18}\text{O}$ values in our surface records (approximately -3.5‰ to -4.0‰ versus PDB) are significantly lower than any previously published data for unquestionably well preserved mid-Cretaceous foraminifera. These $\delta^{18}\text{O}$ minima yield peak temperatures (about 32–33 ± 3 °C) that are significantly warmer than average warm-season SSTs recorded near ODP Site 1052 today (about 26–28 °C) or anywhere else in the modern open ocean.

Our data provide detailed geochemical support for the existence of increased SSTs at low latitude during a past interval of pronounced polar warmth. This result helps to reconcile long-standing discrepancies between palaeoceanographic records, biotic data sets and the predictions of general circulation model (GCM) experiments. If substantiated by data from other low-latitude sites, our

finding would lend support to the widely suggested role of greenhouse-gas forcing as one of the main contributors to global warmth during the mid-Cretaceous. However, the short-term changes in SST indicated by our data contradict the traditional view of relative climate stability during inferred greenhouse periods.

Comparison of our surface temperature records with temperature estimates from thermocline-dwelling fauna allows us to evaluate the history of upper-ocean stratification and to test competing models for OAEs (Fig. 2). Thermocline temperatures show a steady long-term increase during the latest Albian, accompanied by pronounced variations in both SST and the vertical temperature gradient in the upper ocean. Intervals of high SST correspond to periods of strong stratification (temperature gradients across the upper thermocline ≈ 6 °C), whereas cool SSTs correspond to intervals of weak stratification (gradients ≈ 2 °C). The oscillations in stratification culminate in the near-elimination of the vertical $\delta^{18}\text{O}$

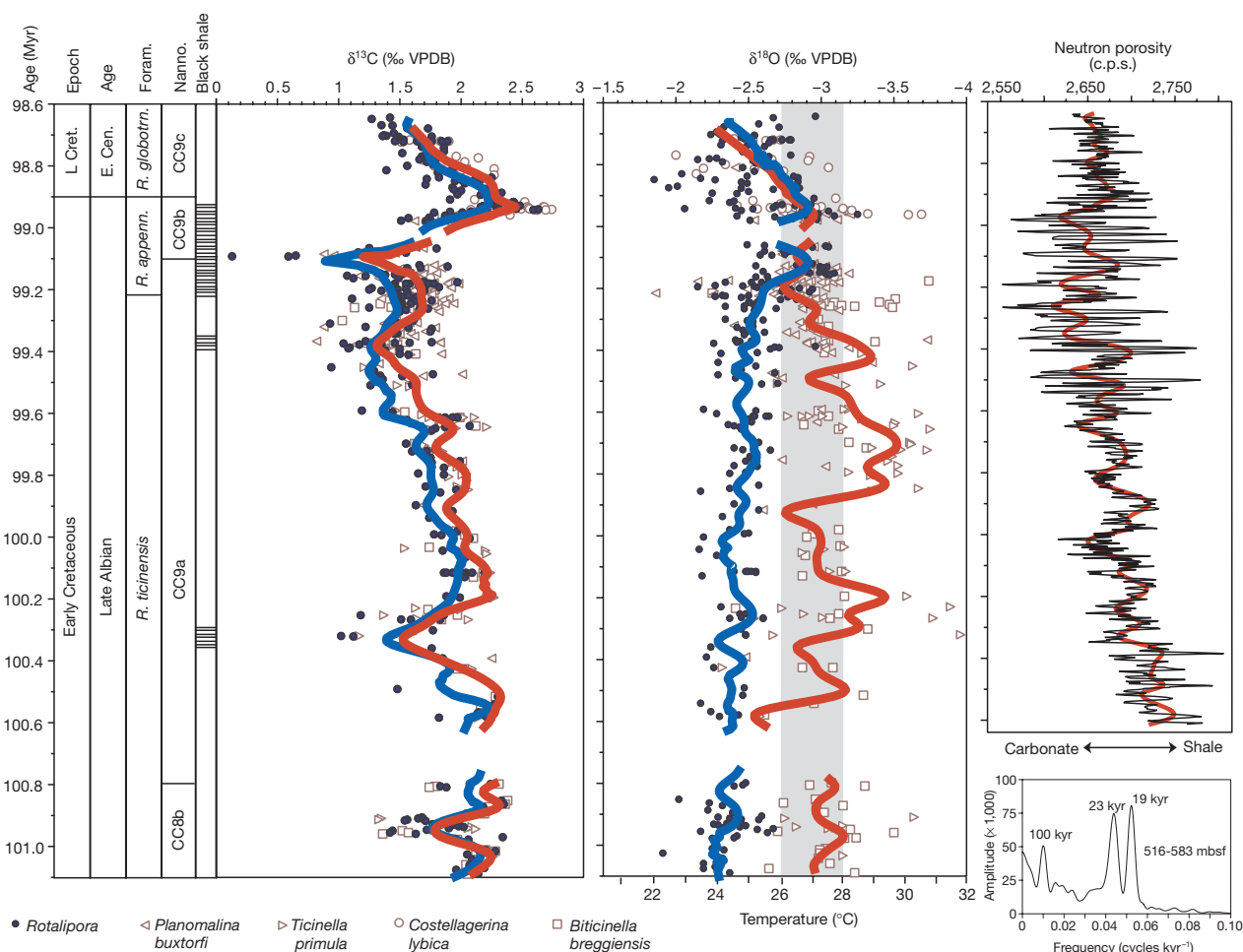


Figure 2 Mid-Cretaceous (late Albian to early Cenomanian) records from ODP Site 1052, Blake nose, western Atlantic. The site (see Fig. 1 for location) is at ~1.3 km water depth. Shown are stable isotope and temperature records, together with the down-hole neutron porosity log. Intervals with black-shale cycles are denoted by black bars. Pronounced cycles from clay-rich carbonate and limestone (low neutron counts) to shale (high neutron counts) in the *R. appenninica* planktonic foraminifera zone are the lithological expression of the latest Albian OAE-1d. The age model was created from the neutron log (see text), assuming that the age of the Albian/Cenomanian boundary is 98.9 Myr (ref. 21). Tuning the log to orbital precession reveals the twin 19- and 23-kyr precession cycles and the 100-kyr eccentricity cycle. Isotope data were generated using pristine planktonic foraminiferal calcite (see text). Solid lines, 100 kyr to infinity filter fit (blue, thermocline; red, surface record). Stippled region, maximum sea surface temperatures in the region of ODP Site 1052 today. All temperatures are calculated using equation (1) of ref. 29, to

which we assign an error estimate (± 3.0 °C) based on the recent discussion of ref. 22. This equation, the outcome of a recent thorough re-evaluation of $\delta^{18}\text{O}$ -palaeothermometry in planktonic foraminiferal calcite, represents the most suitable choice for our non-symbiont bearing fauna¹⁵, and yields very similar temperatures in our dynamic range to those calculated using historically favoured equations for SST estimation from foraminiferal calcite. Following refs 15, 22 and 30, our best estimates for mid-Cretaceous temperatures are those that employ a regional, rather than global, estimate for contemporaneous seawater $\delta^{18}\text{O}$, δw , and we use a conservative value (0.5‰), calculated²⁹ in a recent GCM evaluation of the problem. A slightly higher value for δw (0.78‰) is calculated using an equation based on modern surface ocean δw gradients in the South Atlantic²² and yields correspondingly slightly warmer temperatures (for example, peak SSTs of 33 °C instead of 32 °C). c.p.s., counts per second.

gradient coincident with the onset of OAE-1d and cyclical black-shale development (~99.2 Myr ago) that is reflected in all fauna (Fig. 3). This signal is distinct from previous events (Fig. 2), because it involves not only an abrupt collapse in SSTs but also a marked increase in thermocline temperatures (Fig. 3). Shifts in foraminifera depth habitats are unlikely to explain the collapse in $\delta^{18}\text{O}$ gradient because this would require that surface and thermocline fauna migrated in opposite directions. Thus, the signal appears to reflect a genuine collapse in the upper-ocean thermal gradient, an increase in the depth of the mixed layer and/or a shift in the gradient or seasonality of the thermocline. One possibility is that stratification variability is related to local changes in the influence of proto-western boundary current waters versus subtropical gyre waters. Yet, global anoxia is difficult to explain by local changes in hydrography or upwelling²³. It also seems unlikely that the close correlation between stratification collapse and OAE-1d black shales is coincidental. We suggest that these events involved an increase in deep winter mixing and reduced summer stratification analogous to the modern seasonal cycle in the subtropics. Similar changes in hydrography and productivity are inferred from Neogene records, and have been shown to be orbitally paced²⁴.

The late Albian Breistroffer OAE-1d (~99 Myr ago) differs markedly from its early Albian precursor, the 'Paquier' OAE-1b (~121 Myr ago). The Paquier event is now thought to have been a regional (North Atlantic-western Tethys) 'mega-sapropel' that was associated with intense vertical stratification and did not involve a significant disruption of the global geochemical carbon cycle¹⁷. In contrast, we have shown that the Breistroffer OAE-1d was a global event (Fig. 1) that involved the collapse of upper-ocean stratifica-

tion, at least in the western tropical Atlantic (Fig. 3). Moreover, OAE-1d must represent a significant perturbation to the global carbon cycle, because the main phase of black-shale deposition in Site 1052 is associated with a pronounced (> 2‰) positive $\delta^{13}\text{C}$ excursion that appears correlative to lower-resolution bulk $\delta^{13}\text{C}$ records from European sections^{25,26}. This result provides a valuable chemostratigraphic marker for the boundary between the Lower and Upper Cretaceous.

Positive $\delta^{13}\text{C}$ excursions are now well documented across several major Mesozoic carbon-cycle perturbations (Toarcian, Aptian and Cenomanian/Turonian boundary OAEs)^{5,24–26}, and are thought to reflect the response of the inorganic carbon reservoir to global burial of ^{13}C -depleted organic carbon. In addition, the Aptian and Toarcian OAEs are associated with short-term negative $\delta^{13}\text{C}$ excursions and lags between deposition of organic carbon and the onset of positive $\delta^{13}\text{C}$ excursions^{18–20}. OAE-1d displays similar behaviour, suggesting that some source of ^{13}C -depleted carbon or isotopic fractionation effects act¹⁹ to suppress or override the response of the seawater carbon reservoir to organic-carbon burial in the core of OAE-1d (~99.1 Myr ago, Fig. 2). These similarities imply that common linkages exist between three of the most prominent rapid carbon-cycle perturbations of the past 200 Myr of Earth's history. Our $\delta^{18}\text{O}$ records for OAE-1d provide, to our knowledge, the first quantitative constraints on the physical structure of the upper ocean through any of these perturbations, and are fundamentally inconsistent with 'ocean stagnation' models for their inception (whatever the inferred mechanism for restricted O_2 supply). More generally, our records are consistent with the 'greenhouse' hypothesis for the climates that favoured these extreme carbon-cycle perturbations, but we find no evidence to support the view that such climates were in any way more stable than today. □

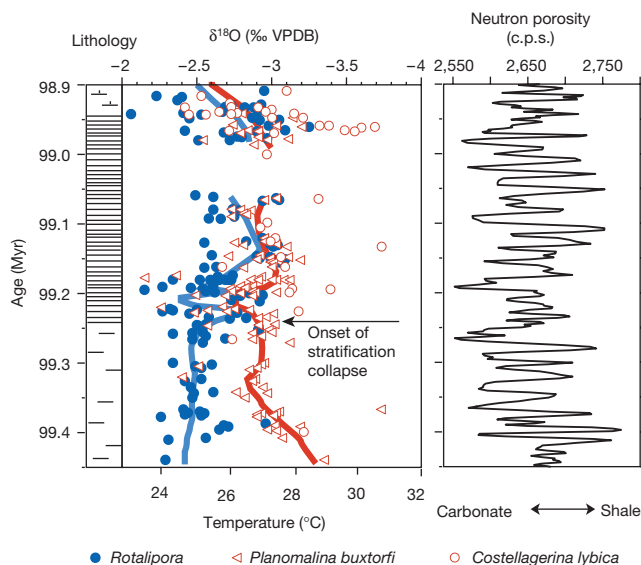


Figure 3 Expanded (550-kyr) $\delta^{18}\text{O}$ and temperature records from ODP Site 1052. These records show the structure of the collapse in upper-ocean stratification that is associated with black-shale deposition during OAE-1d. This event is distinct from previous events (Fig. 2) because it involves both a collapse in SSTs and also a marked increase in thermocline temperatures (and therefore cannot readily be explained in terms of shifting foraminiferal depth habitats, see text). Instead, we suggest that this stratification collapse involved an increase in the depth of the mixed layer and/or shift in the gradient or seasonality of the thermocline, possibly related to some combination of an increase in deep winter mixing and reduced summer stratification (see text). The fact that one or two data points for *Costellagerina lybica* (epipelagic species) plot with outlying low $\delta^{18}\text{O}$ values in the OAE-1d interval strengthens the case that the inferred stratification collapse is a real hydrographic feature that is seasonally modulated. Replacement of the tinnellids by *Planomalina buxtorfi* and *C. lybica* as the epipelagic fauna through time in the record represents global biostratigraphic turnover (rather than local environmental exclusion)^{21,25}. Symbols are the same as in Fig. 2.

Received 29 November 2000; accepted 17 May 2001.

- Fischer, A. G. in *Climate in Earth History* (ed. Berger, W.) 97–104 (National Academy of Sciences, Washington DC, 1982).
- Wilson, P. A., Jenkyns, H. C., Elderfield, H. & Larson, R. L. The paradox of drowned carbonate platforms and the origin of Cretaceous guyots. *Nature* **392**, 889–894 (1998).
- Jenkyns, H. C. Cretaceous anoxic events: from continents to oceans. *J. Geol. Soc. Lond.* **137**, 171–188 (1980).
- Herbert, T. D. & Fischer, A. G. Milankovitch climatic origin of mid-Cretaceous black shale rhythms in central Italy. *Nature* **321**, 739–743 (1986).
- Arthur, M. A., Dean, W. E. & Pratt, L. M. Geochemical and climatic effects of increased marine organic carbon burial at the Cenomanian/Turonian boundary. *Nature* **335**, 714–717 (1988).
- Arthur, M. A., Jenkyns, H. C., Brumsack, H. & Schlanger, S. O. in *Cretaceous Resources, Events and Rhythms* (eds Ginsburg, R. N. & Beaudoin, B.) 75–119 (NATO ASI Series C, 304, Kluwer Academic, Dordrecht, 1990).
- Schlanger, S. O., Jenkyns, H. C. & Premoli-Silva, I. Volcanism and vertical tectonics in the Pacific basin related to global Cretaceous transgressions. *Earth Planet. Sci. Lett.* **52**, 435–449 (1981).
- Arthur, M. A., Dean, W. E. & Schlanger, S. O. in *The Carbon Cycle and Atmospheric CO_2 : Natural Variations Archaean to Present* (eds Sundquist, E. T. & Broecker, W. S.) 504–529 (AGU Monograph 32, American Geophysical Union, Washington DC, 1985).
- Barron, E. J. & Washington, W. M. in *The Carbon Cycle and Atmospheric CO_2 : Natural Variations Archaean to Present* (eds Sundquist, E. T. & Broecker, W. S.) 546–553 (AGU Monograph 32, American Geophysical Union, Washington DC, 1985).
- Manabe, S. & Bryan, K. CO_2 -induced change in a coupled ocean-atmosphere model and its palaeoclimatic implications. *J. Geophys. Res.* **90**, 11689–11708 (1985).
- Crowley, T. J. Past CO_2 changes and tropical sea surface temperatures. *Paleoceanography* **6**, 387–394 (1991).
- Barron, E. in *Effects of Past Global Change on Life* 108–117 (Studies in Geophysics, National Academy Press, Washington DC, 1995).
- Huber, B. T., Hodell, D. A. & Hamilton, C. P. Middle-Late Cretaceous climate of southern high latitudes: stable isotopic evidence for minimal equatorial-to-pole thermal gradients. *Geol. Soc. Am. Bull.* **107**, 1164–1191 (1995).
- Wilson, P. A. & Opdyke, B. N. Equatorial sea surface temperatures for the Maastrichtian revealed through remarkable preservation of metastable carbonate. *Geology* **24**, 555–558 (1996).
- Norris, R. D. & Wilson, P. A. Low latitude sea surface temperatures for the mid-Cretaceous and the evolution of planktonic foraminifera. *Geology* **26**, 823–826 (1998).
- Erbacher, J. & Thurow, J. Influence of oceanic anoxic events on the evolution of mid-Cretaceous radiolaria in the North Atlantic and western Tethys. *Mar. Micropaleontol.* **30**, 139–158 (1997).
- Erbacher, R. D., Huber, B. T., Norris, R. D. & Markey, M. Increased thermohaline stratification as a possible cause for a Cretaceous oceanic anoxic event. *Nature* **409**, 325–327 (2001).
- Erba, E. et al. Integrated stratigraphy of the Cison Apitcore (Southern Alps, Italy): "A reference section" for the Barremian-Aptian interval at low latitudes. *J. For. Res.* **29**, 371–391 (1999).
- Opdyke, B. N., Erba, E. & Larson, R. L. Hot LIPs and mantle methane. *Eos* **80**, F486–F487 (1999).
- Hesselbo, S. P. Massive dissociation of gas hydrate during a Jurassic oceanic anoxic event. *Nature* **406**, 392–395 (2000).

21. Shipboard Scientific Party. Explanatory notes. *Proc. ODP Init. Rep.* **144**, 20 (1995).
22. Crowley, T. J. & Zachos, J. C. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. G. & Wing, S. L.) 50–76 (Cambridge Univ. Press, Cambridge, 2000).
23. Sarmiento, J. L., Herbert, T. D. & Toggweiler, J. R. Causes of anoxia in the world ocean. *Glob. Biogeochem. Cycles* **2**, 115–128 (1988).
24. Wolff, T., Mülitz, S., Rühlemann, C. & Wefer, G. Response of the tropical Atlantic thermocline to late Quaternary trade wind changes. *Palaeoceanography* **14**, 374–383 (1999).
25. Gale, A. S., Kennedy, W. J., Burnett, J. A., Caron, M. & Kidd, B. E. The late Albian to Early Cenomanian succession at Mont Risou near Rosans (Drome, SE France): an integrated study (ammonites, inoceramids, planktonic foraminifera, nannofossils, oxygen and carbon isotopes). *Cret. Res.* **17**, 515–606 (1996).
26. Stoll, H. M. & Schrag, D. P. High-resolution stable isotope records from the Upper Cretaceous rocks of Italy and Spain: Glacial episodes in a greenhouse planet? *Geol. Soc. Am. Bull.* **112**, 308–319 (2000).
27. Smith, A. G., Smith, D. G. & Funnell, B. M. *Atlas of Mesozoic and Cenozoic Coastlines* (Cambridge Univ. Press, Cambridge, 1994).
28. Hofmann, P., Ricken, W., Schwark, R. L. & Leythaeuser, D. Carbon-sulfur-iron relationships in $\delta^{13}\text{C}$ of organic matter from late Albian sedimentary rocks from the North Atlantic Ocean: palaeoceanographic implications. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **163**, 97–113 (2000).
29. Bemis, B. E., Spero, H. J., Bijma, J. & Lea, D. W. Reevaluation of the oxygen isotopic composition of planktonic foraminifera: Experimental results and revised palaeotemperature equations. *Palaeoceanography* **13**, 150–160 (1998).
30. Poulsen, C. J., Barron, E. J., Peterson, W. H. & Wilson, P. A. A reinterpretation of mid-Cretaceous shallow-marine temperatures through model-data comparison. *Palaeoceanography* **14**, 679–697 (1999).

Acknowledgements

We thank M. Cooper, M. Hall and K. Davis for laboratory assistance and manuscript production; M. Arthur and E. Barron for comments on the manuscript; and D. Kroon, H. Elderfield, J. Erbacher, B. Huber, L. Kump, B. Opdyke, C. Poulsen, G. Ravizza, E. Rohling and N. Shackleton for support and discussions. This work was supported by a NERC post-doctoral research fellowship and a UK ODP rapid response grant.

Correspondence and requests for materials should be addressed to P.A.W.
(e-mail: paw1@soc.soton.ac.uk).

Dinosaurian growth patterns and rapid avian growth rates

Gregory M. Erickson*, Kristina Curry Rogers† & Scott A. Yerby‡

* Department of Biological Science and College of Medicine, Florida State University, Tallahassee, Florida 32306-1100, USA

† Science Museum of Minnesota and Macalester College, 120 W. Kellogg Boulevard, St Paul, Minnesota 55102, USA

‡ Department of Biomechanical Engineering, Stanford University, Stanford, California 94305, USA

Did dinosaurs grow in a manner similar to extant reptiles, mammals or birds, or were they unique? Are rapid avian growth rates an innovation unique to birds, or were they inherited from dinosaurian precursors? We quantified growth rates for a group of dinosaurs spanning the phylogenetic and size diversity for the clade and used regression analysis to characterize the results. Here we show that dinosaurs exhibited sigmoidal growth curves similar to those of other vertebrates, but had unique growth rates with respect to body mass. All dinosaurs grew at accelerated rates relative to the primitive condition seen in extant reptiles. Small dinosaurs grew at moderately rapid rates, similar to those of marsupials, but large species attained rates comparable to those of eutherian mammals and precocial birds. Growth in giant sauropods was similar to that of whales of comparable size. Non-avian dinosaurs did not attain rates like those of altricial birds. Avian growth rates were attained in a stepwise fashion after birds diverged from theropod ancestors in the Jurassic period.

Attempts to quantify dinosaurian growth rates have focused on the analysis of bone microstructure, but because dinosaur bones are uniquely composed of tissues with both slow-growing reptilian and rapid-growing avian/mammalian attributes, the results have been inconclusive^{3–11}. In addition, the same bone tissue can form at

substantially different rates^{12,13} and the relationship between localized measures of tissue formation rates and overall body growth are untested for nearly all skeletal elements¹⁴. Broader studies have combined histological age assessments with size indices, but lacked reliable means to evaluate developmental body mass¹⁵.

In most extant animals, mass changes with respect to age show sigmoidal patterns¹⁶. Standardized comparisons of maximum growth rates among phylogenetically and morphologically divergent taxa can be made using values from the exponential stage of development^{17,18}. Analyses of exponential growth among the major groups of extant vertebrates indicate that rates absolutely increase with respect to body mass and that each clade has characteristic rates¹⁷.

To compare whole-body growth rates between the Dinosauria and extant vertebrates, similar quantified data are needed. This requires growth series representing the full range of dinosaur size, shape and phylogeny, and accurate age and mass assessments at all ontogenetic stages. To our knowledge, no study of growth rates has fully met these requirements. However, recent merging of bone histology and scaling principles has provided the requisite tools and data for a single dinosaur (*Psittacosaurus mongoliensis*)¹⁴. At last, the pieces are in place to assess how dinosaurs really grew.

We studied genera of non-avian dinosaurs that span a diversity of phylogeny, size and shape (Fig. 1). We included representatives from most major dinosaurian clades, as well as taxa ranging in size from tiny bipedal theropods to enormous quadrupedal sauropods. Notably, our sample also spans nearly the entire temporal range for non-avian dinosaurs, with the inclusion of taxa from the Early Jurassic to Late Cretaceous periods. We previously generated histological samples and longevity data for two dinosaurs (*Apatosaurus excelsus*⁶ and *Psittacosaurus mongoliensis*¹⁴), and supplemented these data with histological growth series from the literature (*Syntarsus rhodesiensis*³, *Massospondylus carinatus*⁴ and *Maiasaura peeblesorum*⁷). We also generated a histological growth series for *Shuvuuia deserti*, a small, highly derived maniraptoriform theropod (Fig. 1)¹⁹. We assessed age from growth lines in histologically prepared specimens (Fig. 1) and obtained mass estimates through the application of scaling techniques (see Methods). A sigmoidal equation was used to model the growth of each species and least-squares regression analysis was used to fit the curves to these data. Exponential stage growth rates were converted to daily growth rates using the appropriate number of days in the Mesozoic era²⁰. A regression line was fitted to the maximum growth rates for the dinosaurs to enable comparisons with data from the literature for extant vertebrates^{17,18}.

Our analysis revealed that sigmoidal equations (Fig. 2) accurately describe the growth data for the six dinosaurs we tested ($0.885 \leq r^2 \leq 1.0$). Exponential-stage growth rates ranged from 3.4 to 14,460 g day⁻¹, with values positively correlating with increased adult mass (Fig. 3). No taxon substantially deviated from the general dinosaurian trend ($r^2 = 0.96$). The length of the exponential growth stages ranged from about 1 to 6 years (Fig. 2). The onset of somatic maturity occurred between the ages of 3 and 13 years, with values positively correlating with increased body size (Fig. 2).

This research brings us considerably closer to understanding dinosaur biology. It is now possible to quantify dinosaur growth and evaluate the fit of hypotheses regarding maximum dinosaurian growth rates; to address how some dinosaurs attained giant proportions; and to elucidate how and when extremely rapid growth evolved in avian dinosaurs.

Heated debates among palaeontologists and physiologists have revolved around whether the Dinosauria grew like extant reptiles scaled up to giant proportions²¹, like extant birds and/or mammals²², or had growth rates intermediate between these major groups¹¹. The results from our study suggest that none of these models is correct. All dinosaurs grew at rates more rapid than

those of extant reptiles, but as a whole they did not show rates intermediate between reptiles and birds/mammals, nor values equivalent to the latter (Fig. 3). Unexpectedly, our results show that growth in the Dinosauria was unique among major vertebrate groups with members showing rates below, equivalent to, or above typical mammalian/avian rates depending on the size of animal being considered (Fig. 3).

The regression equation for the Dinosauria allows quantitative predictions of dinosaurian growth rates and gross exploration of these data within and beyond the 15,000-fold range of sizes we studied. It is evident that the smallest dinosaurs, such as the dromaeosaurid *Microaptor zhaoianus*²³, with a body mass of ~220 g (on the basis of the mass of *Shuvuuia* and scaling of femoral length), would have grown at ~0.33 g day⁻¹, a rate double that of extant reptiles of comparable adult mass (Fig. 3). Somewhat larger dinosaurs (1–20 kg) grew at rates (1.3–21 g day⁻¹) approximating those of marsupial mammals (Fig. 3). Animals 100–1,000 kg grew at rates (93–786 g day⁻¹) typical of precocial birds, and those 1,500–3,500 kg grew like eutherian mammals (1,144–2,504 g day⁻¹; Fig. 3). Very large sauropods, such as *Apatosaurus* (~25,000 kg),

had growth rates similar to those of whales (absolutely and relatively some of the fastest-growing eutherians). For example, our analysis of a 25,952-kg *Apatosaurus* indicates a growth rate of 14,460 g day⁻¹, compared with 20,700 g day⁻¹ for a 30,000-kg gray whale (*Eschrichtius robustus*)¹⁷. Despite the amazing growth rates for these extinct terrestrial animals, the fastest-growing animal known is still the blue whale (*Balaenoptera musculus*) at 66,000 g day⁻¹ (ref. 17). The largest dinosaurs (for example, *Argentinosaurus*), tentatively estimated at 100,000 kg (ref. 24), are predicted to have grown at absolutely slower rates of 55,638 g day⁻¹.

Despite showing growth rates accelerated from the primitive reptilian character state, non-avian dinosaurs never attained extremely rapid rates like those seen in extant altricial birds (Fig. 3). For example, even the largest sauropods²⁴ would have grown at rates half that of a scaled-up altricial bird (123,025 g day⁻¹; Fig. 3).

So exactly when did dinosaurs attain their unique growth rates? The results for the basal dinosaur taxa (for example, *Syntarsus* and *Massospondylus*) combined with the cosmopolitan distribution of growth rates with respect to mass among our sample suggest that these rates may have been dinosaur apomorphies dating back about

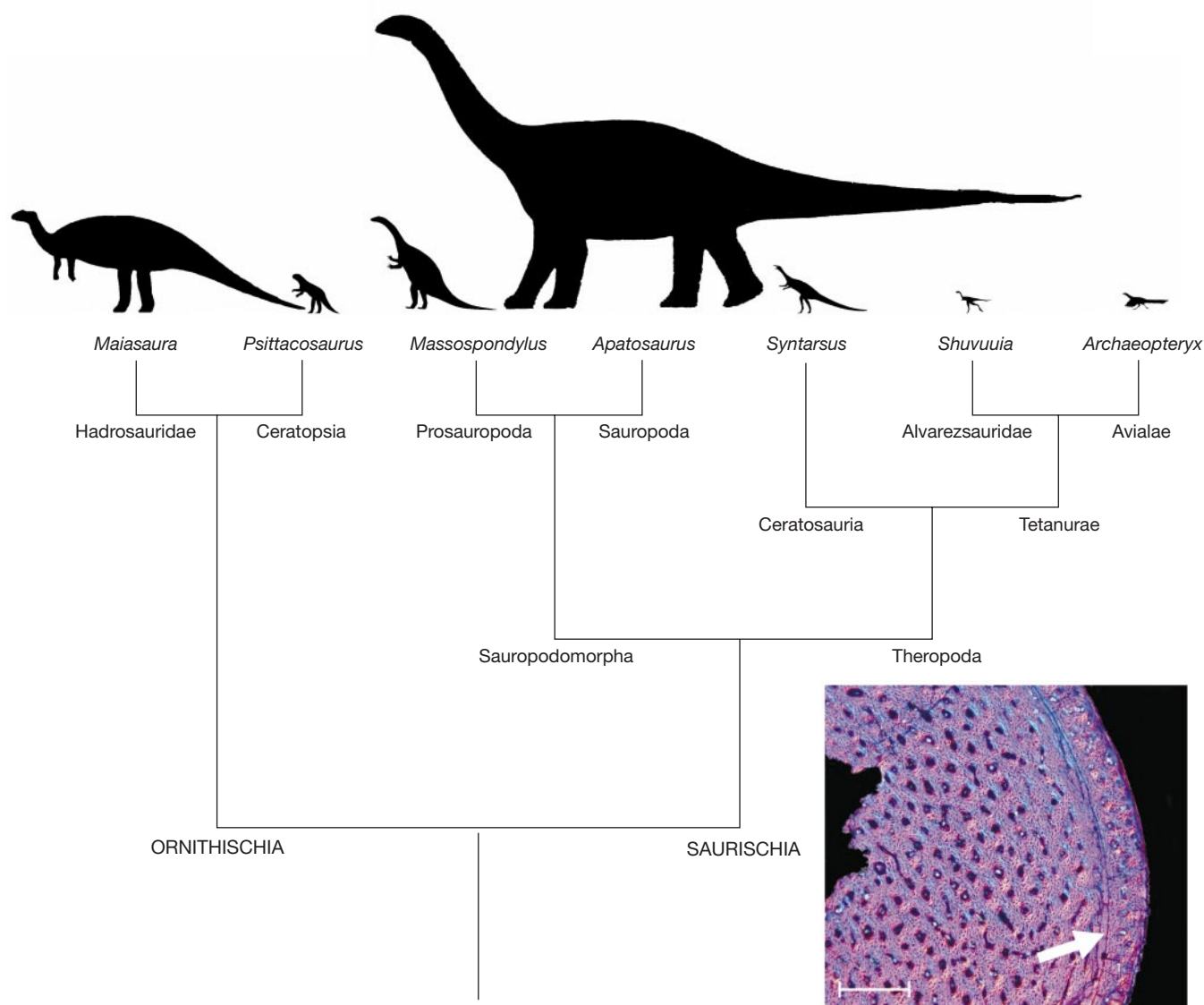


Figure 1 Cladogram for the Dinosauria showing the phylogenetic and size diversity of taxa represented. Six taxa, ranging in size from the 1.7-kg *Shuvuuia deserti* to the 26,000-kg *Apatosaurus excelsus*, were included. Recent phylogenetic analyses show *Shuvuuia*,

once thought to be a bird²⁹, to be a member of a sister clade to Avialae¹⁹. The inset is a histological section of a femur from *S. deserti* (AMNH 100/99) showing a growth line (arrow) used in the ageing of dinosaurs. Scale bar, 0.4 mm.

225 Myr. The first dinosaurs were bipedal forms about 25 kg that are predicted to have had growth rates 6.6 times faster than the primitive character state in reptilian predecessors (25.9 compared with 3.95 g day⁻¹). Whether the derived dinosaurian growth rates are a synapomorphy or if they evolved earlier in a closely related sister taxon (perhaps coincident with erect posture and associated physiological changes) is currently indeterminable.

Birds clearly attained a portion of their elevated growth rates from their dinosaurian ancestry, but when and how did they surpass

the rates of non-avian dinosaurs? Assuming the dinosaurian growth patterns hold true in other outgroups to the Avialae, this event occurred during or after² the evolution of birds. Key evidence for this hypothesis is regression data for small (0.22–20 kg) non-avian dinosaurs (including data for maniraptoriform sister taxa) showing that their growth rates were typically 2–7 times slower than those for similar-sized precocial birds (Fig. 3). Further, stepped evolutionary patterns of avian growth rates later occurred as rates soared fivefold or more above precocial levels and up to 75 times the original

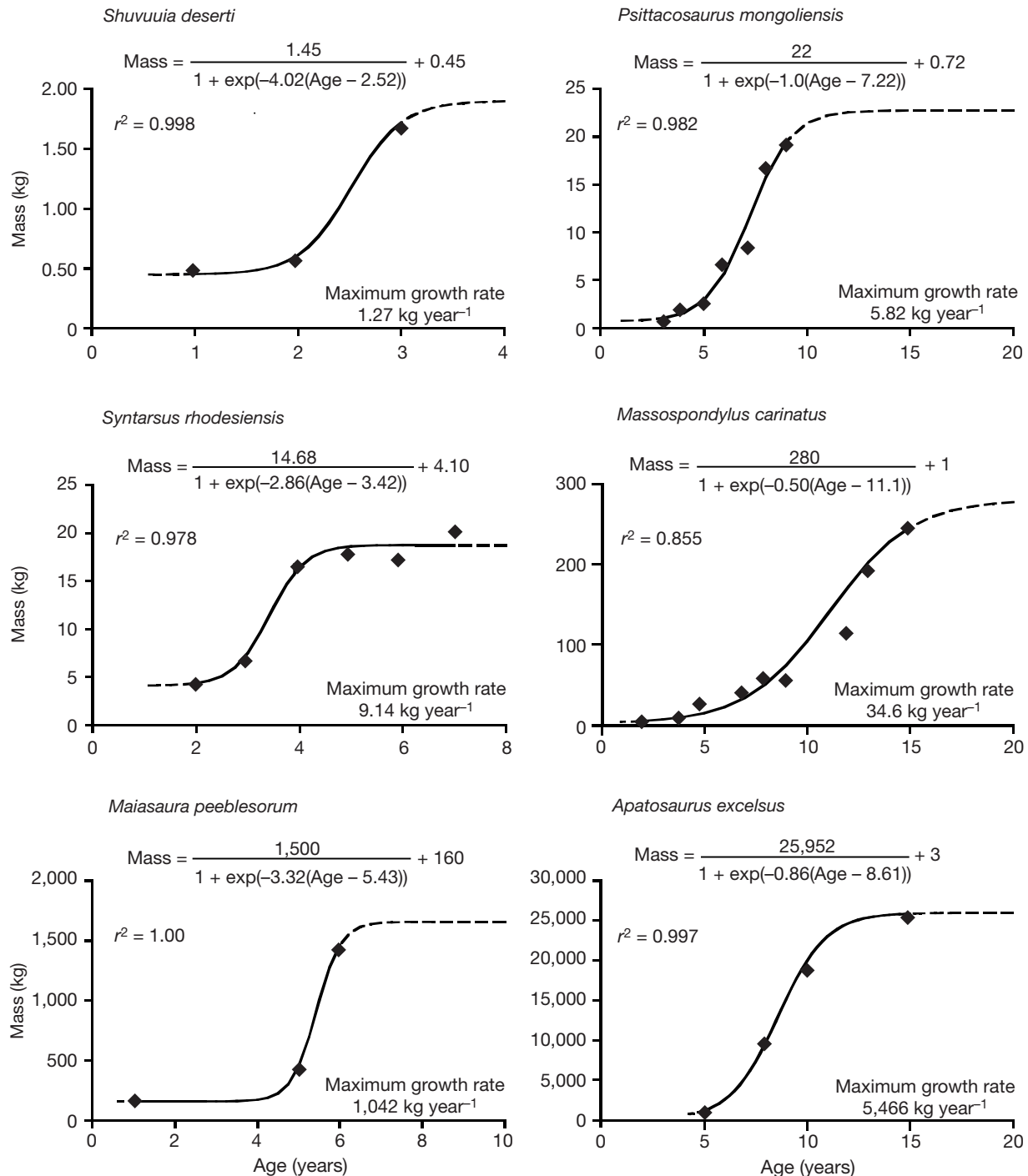


Figure 2 Growth curves for a diversity of dinosaurs. Dinosaur life history was characterized by a slow-growth lag stage during infancy. This was followed by an exponential growth stage midway through development, when maximum growth rates were obtained and most mass was accrued. Development culminated with a stationary

phase, when growth slowed or came to a standstill^{14,16}. Note that the largest animals in the growth series are among the largest specimens known for each taxon and the growth asymptotes were set accordingly¹⁴. The figure for *Psittacosaurus mongoliensis* is modified from Erickson and Tumanova¹⁴.

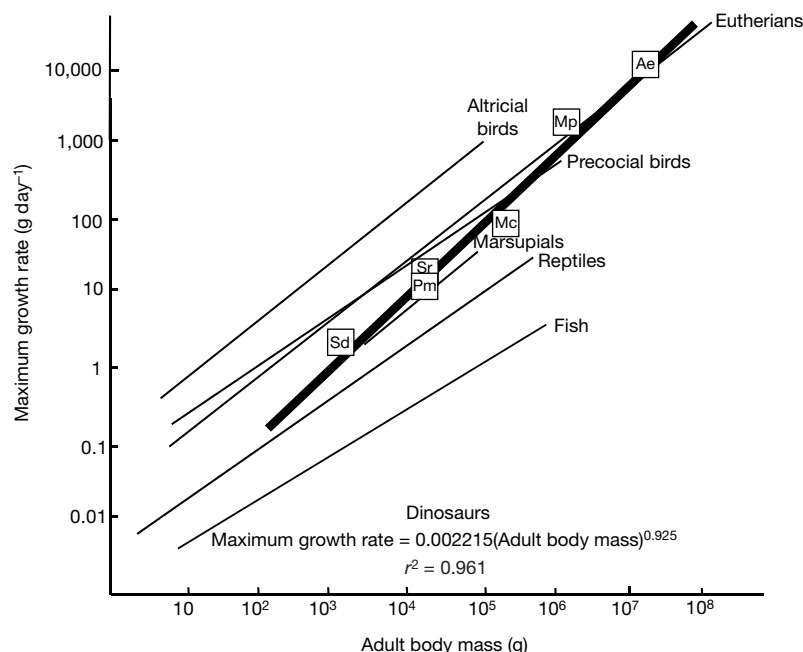


Figure 3 Comparison of exponential-stage growth rates in dinosaurs with typical values for extant vertebrates. Standardized comparisons are made using contrasts between animals of comparable adult mass to diminish signal from differences in shape and negate the effects of size^{17,18}. For example, growth of 20-kg reptiles ($\sim 3.5 \text{ g day}^{-1}$) can be compared to rates for 20-kg altricial birds ($\sim 270 \text{ g day}^{-1}$) to reveal a 75-fold difference in typical growth rates. The dinosaur regression line extends to the bounds of the known size range for the clade ($\sim 0.2\text{--}100,000 \text{ kg}$). Letters represent growth rates: Sd, *Shuvuuia deserti* (3.4 g day^{-1}); Pm, *Psittacosaurus mongoliensis* (12.5 g day^{-1});

Sr, *Syntarsus rhodesiensis* (23.9 g day^{-1}); Mc, *Massospondylus carinatus* (90.3 g day^{-1}); Mp, *Maiasaura peeblesorum* ($2,793 \text{ g day}^{-1}$); Ae, *Apatosaurus excelsus* ($14,460 \text{ g day}^{-1}$). The steep slope for the Dinosauria is unique among major vertebrate clades but is not unprecedented for minor vertebrate clades spanning smaller ranges of body size¹⁷. Note that there is some overlap among groupings for individual taxa (not shown). For example, primates are extremely slow-growing eutherians, showing rates similar to those of marsupials¹⁷. Data for extant groups and graphics are modified from Case¹⁷ and Calder¹⁸.

reptilian condition in association with the evolution of altriciality^{17,18} (Fig. 3). □

Methods

Age assessment

Growth rings were counted in histological sections of each specimen²⁵. We accounted for loss of growth rings due to medullar expansion with increased age by sequentially superimposing subadult specimens upon those from larger individuals. The annual periodicity of these growth lines is established on the grounds of phylogenetic parsimony and tissue formation rates consistent with those for extant taxa^{14,25}.

Assessments of body mass

For all specimens except *Syntarsus*³ and *Massospondylus*⁴, body mass estimates were required. These were assessed using long-bone diaphyseal circumferences and regression equations from ref. 26. As these equations are invalid throughout development (that is, subadult animals of large taxa do not scale in proportion to adults of smaller taxa during ontogeny), we used developmental mass extrapolation¹⁴, a scaling principle like that developed in the middle of the last century²⁷. The accuracy of this methodology was tested for the purposes of the present study on human stature data for approximately 100 individuals²⁷ and on 34 wild alligators studied periodically throughout 8 years of development (G.M.E. and A. R. Woodward, unpublished data). Predictions of exponential growth rates were within 5% in each case (4.4% and 2.8%, respectively). This small degree of uncertainty and the general conformity ($\pm 50\%$) of our estimates of adult mass to other recent methodologies²⁸ suggest that the overall conclusions of this research are robust.

Received 10 May; accepted 6 July 2001.

- Ricqlès, A. de. In *A Cold Look at the Warm-Blooded Dinosaurs* (eds Thomas, D. K. & Olson, E. C.) 103–139 (Westview, Boulder, 1983).
- Chinsamy, A., Chiappe, L. M. & Dodson, P. Growth rings in Mesozoic birds. *Nature* **368**, 196–197 (1994).
- Chinsamy, A. Physiological implications of the bone histology of *Syntarsus rhodesiensis* (Saurischia: Theropoda). *Palaeontol. Afr.* **27**, 77–82 (1990).
- Chinsamy, A. Bone histology and growth trajectory of the prosauropod dinosaur *Massospondylus carinatus* Owen. *Mod. Geol.* **18**, 319–329 (1993).
- Varricchio, D. J. Bone microstructure of the Upper Cretaceous theropod dinosaur *Troodon formosus*. *J. Vertebr. Paleontol.* **13**, 99–104 (1993).
- Curry, K. A. Ontogenetic histology of *Apatosaurus* (Dinosauria: Sauropoda): new insights on growth rates and longevity. *J. Vertebr. Paleontol.* **19**, 654–665 (1999).
- Horner, J. R., Ricqlès, A. de & Padian, K. Long bone histology of the hadrosaurid dinosaur *Maiasaura*

- peeblesorum*: growth dynamics and physiology based on an ontogenetic series of skeletal elements. *J. Vertebr. Paleontol.* **20**, 115–129 (2000).
- Sander, P. M. Long bone histology of the Tendaguru sauropods: implications for growth and biology. *Paleobiology* **26**, 466–488 (2000).
- Ricqlès, A. de. Zonal “growth rings” in dinosaurs. *Mod. Geol.* **15**, 19–48 (1990).
- Rimblot-Baly, F., Ricqlès, A. de & Zylberberg, L. Analyse paléohistologique d’une série de croissance partielle chez *Lapparentosaurus madagascariensis* (Jurassique Moyen): essai sur la dynamique de croissance d’un dinosaure sauropode. *Ann. Paléontol.* **81**, 49–86 (1995).
- Reid, R. E. H. In *The Complete Dinosaur* (eds Farlow, J. O. & Brett-Surman, M. K.) 449–473 (Indiana Univ. Press, Bloomington, 1997).
- Ricqlès, A. de, Meunier, F. J., Castanet, J. & Francillon-Vieillot, H. In *Bone Vol. 3* (ed. Hall, B. K.) 1–78 (CRC, Boca Raton, 1991).
- Castanet, J., Curry Rogers, K., Cubo, J. & Boisard, J. J. Quantification of periosteal osteogenesis in ostrich and emu: implications for assessing growth in dinosaurs. *C.R. Acad. Sci. III* **323**, 543–550 (2000).
- Erickson, G. M. & Tumanova, T. A. Growth curve of *Psittacosaurus mongoliensis* Osborn (Ceratopsia: Psittacosauridae) inferred from long bone histology. *Zool. J. Linn. Soc.* **130**, 551–566 (2000).
- Varricchio, D. J. In *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 282–288 (Academic, San Diego, 1997).
- Sussman, M. *Growth and Development* (Prentice-Hall, New Jersey, 1964).
- Case, T. J. On the evolution and adaptive significance of postnatal growth rates in the terrestrial vertebrates. *Q. Rev. Biol.* **53**, 243–282 (1978).
- Calder, W. A. III. *Size, Function, and Life History* (Harvard Univ. Press, Cambridge, Massachusetts, 1984).
- Norell, M. A., Clark, J. M. & Makovicky, P. J. In *New Perspectives on the Origin and Early Evolution of Birds: Proc. Int. Symp. in Honor of John H. Ostrom* (eds Gauthier, J. & Gall, L.) (Special Publ. Peabody Mus. Nat. Hist., New Haven, in the press).
- Wells, J. W. Coral growth and geochronometry. *Nature* **197**, 948–950 (1963).
- Case, T. J. Speculations on the growth rate and reproduction of some dinosaurs. *Paleobiology* **4**, 320–328 (1978).
- Bakker, R. T. *The Dinosaur Heresies* (Kenningson, New York, 1986).
- Xu, X., Zhou, Z. & Wang, X. The smallest known non-avian theropod dinosaur. *Nature* **408**, 705–708 (2000).
- Appenzeller, T. Argentine dinos vie for heavyweight title. *Science* **266**, 1805 (1994).
- Castanet, J. & Smirina, E. Introduction to the skeletochronological method in amphibians and reptiles. *Ann. Sci. Nat. Zool.* **11**, 191–196 (1990).
- Anderson, J. F., Hall-Martin, A. & Russell, D. A. Long-bone circumference and weight in mammals, birds and dinosaurs. *J. Zool.* **A207**, 53–61 (1985).
- Thompson, D. W. *On Growth and Form* (Cambridge Univ. Press, New York, 1943).
- Seebacher, F. A new method to calculate allometric length-mass relationships of dinosaurs. *J. Vertebr. Paleontol.* **21**, 51–60 (2001).
- Chiappe, L. M., Norell, M. A. & Clark, J. A. The skull of a relative of the stem-group bird *Mononychus*. *Nature* **392**, 275–278 (1998).

Acknowledgements

We thank J. Horner and the staff of the Museum of the Rockies; M. Norell and staff of the American Museum of Natural History; A. Woodward and the Florida Fish and Wildlife Conservation Commission; and T. Tumanova, B. Breithaupt, P. Makovicky, M. Sander, D. Varricchio, E. Creech, R. Rogers, B. Erickson and K. Womble. This research was funded by the College of Arts and Sciences and Department of Biological Science of Florida State University.

Correspondence and requests for materials should be addressed to G.M.E. (e-mail: gerickson@bio.fsu.edu).

Mealybug β -proteobacterial endosymbionts contain γ -proteobacterial symbionts

Carol D. von Dohlen, Shawn Kohler*, Skylar T. Alsop & William R. McManus

Department of Biology, 5305 Old Main Hill, Utah State University, Logan, Utah 84322, USA

Some insects have cultivated intimate relationships with mutualistic bacteria since their early evolutionary history. Most ancient 'primary' endosymbionts live within the cytoplasm of large, polyploid host cells of a specialized organ (bacteriome)¹. Within their large, ovoid bacteriomes, mealybugs (Pseudococcidae) package the intracellular endosymbionts into 'mucus-filled' spheres, which surround the host cell nucleus and occupy most of the cytoplasm². The genesis of symbiotic spheres has not been determined, and they are structurally unlike eukaryotic cell vesicles. Recent molecular phylogenetic and fluorescent *in situ* hybridization (FISH) studies suggested that two unrelated bacterial species may share individual host cells^{3,4}, and that bacteria within spheres comprise these two species⁵. Here we show that mealybug host cells do indeed harbour both β - and γ -subdivision Proteobacteria, but they are not co-inhabitants of the spheres. Rather, we show that the symbiotic spheres themselves are β -proteobacterial cells. Thus, γ -Proteobacteria live symbiotically inside β -Proteobacteria. This is the first report, to our knowledge, of an intracellular symbiosis involving two species of bacteria.

Most members of the large Hemipteran suborder Sternorrhyncha (aphids, whiteflies, psyllids, scales and mealybugs) feed on nutrient-deficient plant sap, and appear to fortify their diet with the metabolic products of mutualistic bacteria⁶. Mutualisms between insects and primary endosymbionts may date to the origins of the host families or superfamilies (100–250 Myr ago)^{7,8}. More recently acquired secondary endosymbionts are sometimes harboured in unspecialized syncytial or epithelial cells of the bacteriome^{1,8}, except in whiteflies, where two or more bacterial forms occupy the specialized host cells⁹. Endosymbionts are transferred vertically from maternal host cells to eggs *in vivo* in a highly organized process that reflects their antiquity¹. On the basis of molecular phylogenetic analyses of 16S ribosomal DNA, most insect endosymbionts have been identified as γ -subdivision Proteobacteria (purple bacteria), related to enterics such as *E. coli* and *Salmonella*¹⁰. Thus, the first report of mealybug endosymbionts as β -subdivision Proteobacteria (related to *Burkholderia*) was unusual⁴. *In situ*

hybridization of a γ -proteobacterial sequence to the bacteriome³ conflicted with that report⁴, but later seemed to be reconciled by the localization of both β - and γ -proteobacterial sequences to host cells⁵.

The peculiar mucoidal compartments within mealybug host cells have long puzzled researchers. Early studies detected RNA, ribosome-like granules, glycoproteins, microtubules and crystalline bodies in the mucus-like material of the spheres, but no cellular organelles¹¹. Protein synthesis within spheres (with a sphere and its bacteria treated as a unit) was independent from that of host cells¹². The ingestion of antibiotics by mealybugs, or exposure to high temperatures, resulted in degeneration of both bacteria and spheres within a few days; the insects died soon after¹³. Early electron microscopy revealed three membrane bilayers surrounding symbiotic spheres¹⁴; we note that this would be an unprecedented structure for eukaryotic cell vesicles.

We used polymerase chain reaction (PCR) and cloning to amplify and separate 16S rDNAs from bacteria present in bacteriome tissue of the citrus mealybug, *Planococcus citri* (Risso). Sequencing revealed similar numbers of clones from a β -subdivision proteobacterium (six) and a γ -subdivision proteobacterium (four). Quantitative PCR also suggested that numbers of both bacterial 16S gene copies in the bacteriome were similar—that is, neither result was due to a stray contaminant (amplification of both 16S genes most closely matched the 50,000 copy plasmid standards). The β -proteobacterial sequence was identical to a published bacterial sequence from *P. citri* (GenBank accession number M68890.2; originally as *Pseudococcus maritimus*)⁴. The γ -proteobacterial sequence was 93–94% identical to a tsetse fly secondary endosymbiont, a weevil primary endosymbiont, and several *Erwinia* spp.

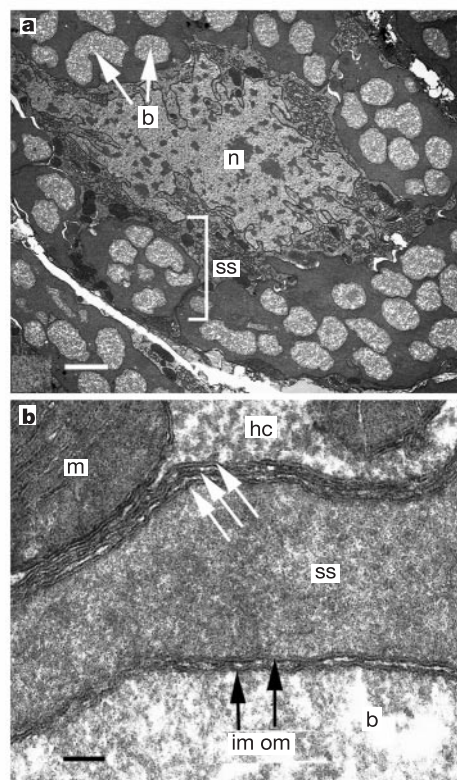


Figure 1 Transmission electron micrographs showing structure within specialized host cells. **a**, Host cell with central nucleus surrounded by seven spheres containing bacteria. Scale bar is 2.33 μ m. **b**, High magnification of two membranes of a bacterium (black arrows) and three membranes of a symbiotic sphere (white arrows). Scale bar is 0.0706 μ m. b, bacteria; hc, host cell cytoplasm; im, inner membrane; m, mitochondrion; n, nucleus; om, outer membrane; ss, symbiotic sphere.

* Present address: LDN, NICHD, National Institutes of Health, MSC 4480, 9000 Rockville Pike, Bethesda, Maryland 20892, USA.

Using transmission electron microscopy, we confirmed the organization of mealybug host cells and the membrane structure of spheres and resident endosymbionts. Sections from dissected bacteriomes exhibited host cells with large, central nuclei surrounded by spheres filled with bacteria (Fig. 1a). Bacteria were of consistent size and density, with no distinctive second form. Bacteria within spheres were surrounded by two membrane bilayers, while spheres were surrounded by three membrane bilayers (Fig. 1b).

We used FISH visualized by laser-scanning confocal microscopy to localize the signal from labelled probes hybridized to bacterial

rRNAs. FISH with a general eubacterial 16S rRNA probe resulted in a strong signal localized to the bacteriome of immature nymphs (Fig. 2a). Bacteria within spheres stained most brightly, but the matrix of the spheres also stained to a lesser degree (Fig. 2b; we note how spheres are outlined by fluorescence). FISH performed with the gamma 16S-specific probe resulted in hybridization only to the bacteria within spheres (Fig. 2c). FISH undertaken with the beta 16S-specific probe resulted in hybridization to the sphere matrix surrounding the γ -probed bacteria (Fig. 2d). FISH performed with both probes in the same reaction revealed a composite pattern identical to those with single probes: fluorescence from the γ probe indicated bacteria within spheres, and fluorescence from the β probe highlighted the sphere matrix surrounding γ -probed bacteria (Fig. 2e–g). The pattern of probing was consistent with the ultra-structure of spheres and position of endosymbionts as revealed by electron microscopy. Staining with propidium iodide (a fluorescent nucleic-acid-intercalating agent) also produced a strong signal in bacteria within spheres, and a weaker signal in the sphere matrix (see below, Fig. 3a).

We also documented stages in the transmission of symbionts to offspring¹. Host cells of gravid females break down, and freed spheres cluster around the nutrient plasma cord connecting nurse cells to an oocyte's anterior pole (Fig. 3a). The spheres eventually penetrate the oocyte (Fig. 3b). At a later stage following the initiation of embryogenesis, spheres drift toward the embryo's host cells, which simultaneously move from the middle of the embryo to engulf them¹. FISH performed with both γ - and β -specific probes at these stages of transmission revealed the same probing pattern as in maternal host cells, where the γ probe highlighted bacteria in spheres and the β probe highlighted the surrounding matrix of the spheres (Fig. 3c, d).

We conclude that the symbiotic spheres are cells of β -Proteobacteria, and thus that they harbour cells of γ -Proteobacteria within their cytoplasm. This conclusion is tenable for several reasons. First, the sizes of both spheres (~ 10 – $20\ \mu\text{M}$) and bacteria within them

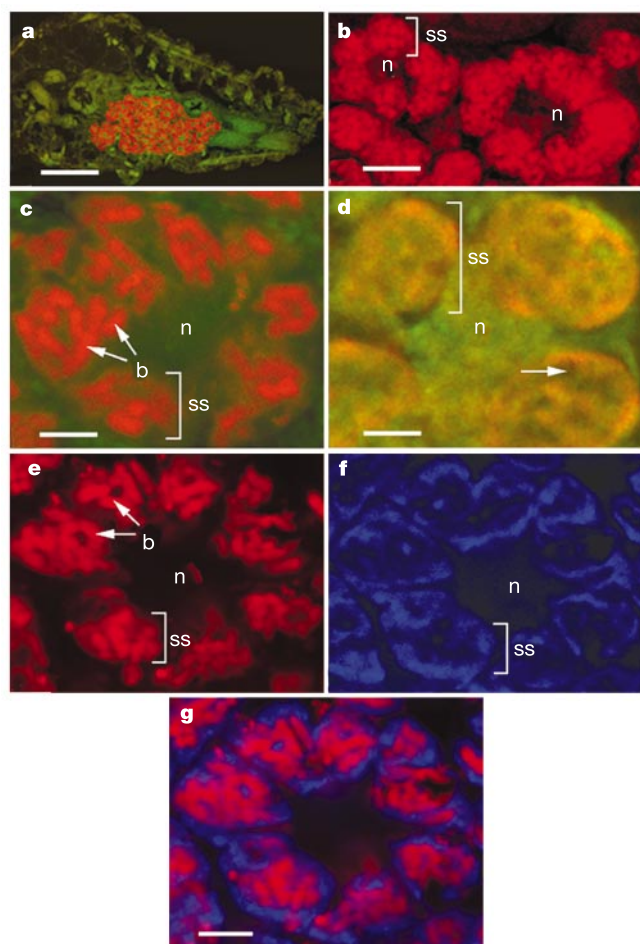


Figure 2 Laser-scanning confocal microscope images of FISH results with specific eubacterial probes on mealybug sections. **a**, Hybridization with the general eubacterial 16S probe (1507–) labelled with AlexaFluor 568: autofluorescence (green) highlights the mealybug body; hybridized probe (red) highlights the bacteriome. Scale bar is 0.42 mm. **b**, Hybridization as in **a**: high magnification showing two entire host cells, where hybridized probe (in red) highlights the symbiotic spheres and resident bacteria. Scale bar is 10.5 μm . **c**, Hybridization with the γ -proteobacterial probe g630 (AlexaFluor 568 label) showing a single host cell: bacteria inside spheres are highlighted in red and autofluorescence is green. Scale bar is 4.5 μm . **d**, Hybridization with the β -proteobacterial probe b91 (AlexaFluor 568 label) showing a single host cell: the matrix of four spheres is highlighted in red and autofluorescence is green. Unlabelled arrow points to the position of a resident bacterium. Scale bar is 4.7 μm . **e**, Hybridization with mixtures of γ - and β -proteobacterial-specific probes simultaneously, showing fluorescence from the γ mix (g80, g458, g630, g1273; AlexaFluor 568 label) in a single host cell; only bacteria within spheres are highlighted (in red). **f**, Same reaction and host cell as in **e**, showing fluorescence from the β mix (b91, b162, b886, B1273; BODIPY 650/665 label); only the matrix of spheres is highlighted (in blue; note how the edges of spheres are outlined). **g**, Same reaction and host cell as in **e**, combining fluorescence from both the γ mix (red) and β mix (blue). Scale bar is 4.5 μm . n, nucleus; b, γ -proteobacteria; ss, symbiotic sphere.

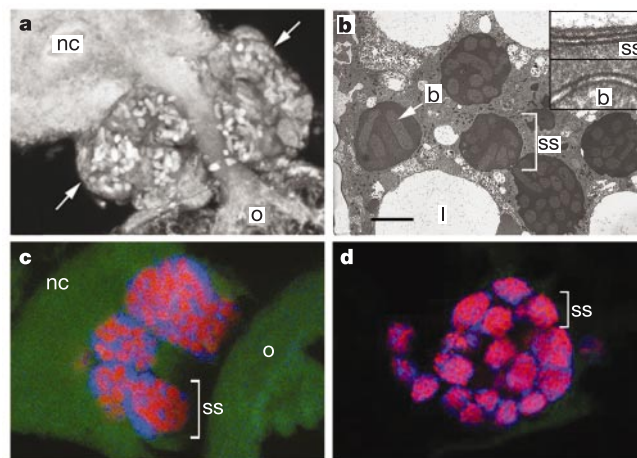


Figure 3 Stages in the transmission of symbiotic spheres to oocytes. **a**, Laser-scanning confocal micrograph of a section from a gravid female mealybug stained with propidium iodide. Arrows point to spheres, containing bacteria, surrounding the nutrient cord connecting nurse cells to oocyte. **b**, Transmission electron micrograph of a mealybug oocyte, showing spheres containing bacteria, which have infected the egg, but have not yet entered host cells. Insets show three membranes of the spheres and two membranes of the γ -Proteobacteria. Scale bar is 6.36 μm . **c**, Confocal micrograph of a hybridization with g630 (AlexaFluor 568 label) and b91 (BODIPY 650/665 label) probes in gravid females. Spheres (blue) containing γ -Proteobacteria (red) are clustered around the nutrient cord connecting nurse cells to oocyte, as in **a**. **d**, Same reaction conditions as in **c**, showing a cluster of spheres (blue) containing γ -Proteobacteria (red) that have penetrated an oocyte but have not yet entered host cells, as in **b**. b, γ -Proteobacteria; l, lipid; n, nucleus; nc, nurse cell; o, oocyte; ss, symbiotic sphere.

(~3–6 μ M) fall within the range of known eubacterial cells; in addition, β -Proteobacteria are known to be endosymbionts of other eukaryotes¹⁵. Second, optical sections by laser-scanning confocal microscopy through the centre of symbiotic spheres indicated signals from hybridized γ - and β -proteobacterial rRNAs in positions corresponding to the ultrastructure of double-membrane-bound bacteria within spheres, and to the matrix of the spheres, respectively (compare Fig. 1a to Fig. 2g and Fig. 3b to d). Third, the three membrane bilayers surrounding symbiotic spheres constitute the same structure found in many insect primary endosymbionts, where the two membranes of the bacterial cell are enclosed by a third, host-derived membrane¹⁶. No known metazoan organelle or vesicle has such a membrane structure. Finally, symbiotic spheres are the units transferred vertically from mother to offspring¹. If the spheres are β -Proteobacteria as we propose, then this mode of transmission is equivalent to that found in most other insects, where transfer is accomplished by individual endosymbiont cells¹.

This is the first report, to our knowledge, of an extant, intracellular bacterial–bacterial symbiosis. We use the term symbiosis because it implies persistent contact between partners in a long-term association¹⁷. In mealybugs, the association appears to be stable over the entire insect life cycle: in many whole-mount, hybridized sections of immatures and gravid adults (including eggs), the γ -Proteobacteria were never detected outside the symbiotic spheres (β -Proteobacteria), nor were empty spheres found. Bacteria are housed within symbiotic spheres in at least six mealybug genera; in species of eight other genera, bacteria live freely within the host cell¹. The identity of these latter bacteria is unknown. Molecular phylogenetic analyses, however, suggest that β -Proteobacteria are the primary endosymbionts of mealybugs with the γ – β symbiosis: their 16S sequences compose a unique lineage^{4,5}, implying a single, ancient infection followed by parallel speciation of endosymbionts and hosts¹⁸. However, γ -proteobacterial endosymbionts of mealybugs appear not to share a common ancestor⁵. Thus, the unique endobacterial symbiosis has either originated more than once during mealybug diversification, or horizontal transfers or replacements of an original γ -proteobacterium have occurred.

The nature of this unusual symbiosis remains to be explored. One possibility is that γ -Proteobacteria are essentially parasites, using the β -Proteobacteria for protection and transport to the next host generation. Considering the emerging evolutionary dynamics of endosymbiosis, however, we hypothesize an association approaching mutualism. Similar to organelles, long-term intracellular symbionts have been altered genetically compared to free-living bacteria¹⁹. The accumulation of slightly deleterious alleles by genetic drift²⁰, resulting in elevated substitution rates and nucleotide bias^{20,21}, has affected rRNA stability^{22,23}, adaptive codon usage^{21,24}, and genome size²⁵. Selection at the insect level, however, may have favoured the evolution of compensatory mechanisms in endosymbionts^{20,26}. Additional compensation might conceivably occur through genetic exchange with other intracellular associates²⁷. We hypothesize that the internalization of γ -Proteobacteria by β -Proteobacteria may be a form of compensation, facilitating the exchange of genes and gene products to slow or reverse genetic degradation. This might explain why the 16S rDNAs of mealybug β -proteobacterial endosymbionts atypically show no nucleotide bias²⁰. □

Methods

Gene amplification, cloning and sequencing

Whole bacteriomes were dissected from live, subadult citrus mealybugs, and DNA was extracted by a protein salting-out method²⁸. Bacterial 16S rDNA was amplified by PCR with general eubacterial primers (10+ and 1507–)² and the products were cloned using the TOPO-TA cloning kit (Invitrogen). Ten clones were sequenced with 10+, and one each of the γ - and β -proteobacterial clones were sequenced completely with internal primers⁷. Similarities to published sequences were determined by the BLAST program of the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>).

Quantitative PCR

To estimate the relative numbers of γ - and β -proteobacterial 16S templates in bacteriome DNA preparations, quantitative PCR was performed as recommended (see <http://www.promega.com/uk/quantpcr.htm>). 16S rRNA genes were amplified from bacteriome DNA separately with 10+ and a γ - or β -proteobacterial-specific primer, as well as from plasmids containing whole 16S inserts (using starting templates of 3,000, 25,000 and 50,000 copies). Amplification was estimated visually on an ethidium bromide-stained agarose gel.

Electron microscopy

Mealybug abdomens, dissected bacteriomes and eggs were fixed in glutaraldehyde and cacodylate buffer following a standard microwave protocol (see http://www.tedpella.com/Micro_html/vacpro.htm). Samples were embedded, sectioned, stained, and then imaged with a Zeiss 902 transmission electron microscope.

Fluorescent *in situ* hybridization

Oligonucleotide probes unique to the γ - and β -proteobacterial 16S rDNAs amplified from mealybugs were designed from an alignment of those sequences plus published 16S rDNA sequences from other insect endosymbionts and free-living Proteobacteria (retrieved from GenBank). Probes were used in hybridization reactions either singly or in mixtures. β -proteobacterial-specific probes were: b91, GCCTTAGCCGTCGCTCCGTAC; b162, TTTCGCCAGTCTATCCCC; b886, TCAGGCGGTCGACTTCAT; and b1273, CCTCTCTCGAGTTGGCGA. γ -proteobacterial-specific probes were: g80, TTTTTCGGTTGCCGCTCAACT; g458, ATTGCCTTCTCCCTACTG; g630, CGAGA CTTAGCCCTATCAGTTTC; and g1273, ACTCTCGCGAGCTTGCTT. The PCR primer 1507– was employed as a general, eubacterial probe. Probes were labelled with either ChromaTide AlexaFluor 568-5-dUTP or ChromaTide BODIPY 650/665-14-dUTP fluorophores (Molecular Probes, Inc.). Whole mealybugs preserved in acetone were fixed, sucrose-infused, frozen in tissue-mounting medium and sectioned with a cryostat²⁹. Sections were probed using standard methods³⁰, with excess unlabelled probe and RNase controls, and results visualized with a BioRad MRC 1024 laser-scanning confocal microscope.

Voucher specimens and GenBank accession numbers

Voucher specimens of *P. citri* (accession numbers 9901–9905, collected April and October 1999 from the Department of Biology greenhouse, Utah State University) were deposited in the Utah State University Insect Collection. Sequences of endosymbiont rDNAs were deposited in GenBank under accessions AF322016–7.

Received 29 November 2000; accepted 25 May 2001.

- Buchner, P. *Endosymbiosis of Animals with Plant Microorganisms* (Wiley, New York, 1965).
- Fink, R. Morphologische und physiologische Untersuchungen an den intrazellulären Symbionten von *Pseudococcus citri* Risso. *Z. Morph. Ökol. Tiere* **41**, 78–146 (1962).
- Kantheti, P., Jayarama, K. S. & Sharat Chandra, H. Developmental analysis of a female-specific 16S rRNA gene from mycetome-associated endosymbionts of a mealybug, *Planococcus lilacinus*. *Insect Biochem. Molecul. Biol.* **26**, 997–1009 (1996).
- Munson, M. A., Baumann, P. & Moran, N. A. Phylogenetic relationships of the endosymbionts of mealybugs (Homoptera: Pseudococcidae) based on 16S rDNA sequences. *Mol. Phylog. Evol.* **1**, 26–30 (1992).
- Fukatsu, T. & Nikoh, N. Endosymbiotic microbiota of the bamboo pseudococcid *Antonina crawii* (Insecta, Homoptera). *Appl. Environ. Microbiol.* **66**, 643–650 (2000).
- Houk, E. J. & Griffiths, G. W. Intracellular symbionts of the Homoptera. *Annu. Rev. Entomol.* **25**, 161–187 (1980).
- Spaulding, A. W. & von Dohlen, C. D. Phylogenetic characterization and molecular evolution of bacterial endosymbionts in psyllids (Hemiptera: Sternorrhyncha). *Mol. Biol. Evol.* **15**, 1506–1513 (1998).
- Moran, N. A. & Telang, A. Bacteriocyte-associated symbionts of insects. *Bioscience* **48**, 295–304 (1998).
- Costa, H. S., Westcott, D. M., Ullman, D. E. & Johnson, M. W. Ultrastructure of the endosymbionts of the whitefly, *Bemisia tabaci* and *Trialeurodes vaporariorum*. *Protoplasma* **176**, 106–115 (1993).
- Baumann, P., Moran, N. A. & Baumann, L. In *The Prokaryotes, A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications* (ed. Dworkin, M.) (Springer, New York, 2000).
- Tremblay, E. In *Insect Endocytobiosis: Morphology, Physiology, Genetics, Evolution* (eds Schwemmler, W. & Gassner, G.) 145–173 (CRC Press, Boca Raton, 1989).
- Louis, C. & Kuhl, G. Synthèse *in situ* de protéines chez les symbiotes et chez les mycétocytes de *Pseudococcus obscurus* (Coccidae-Pseudococcinae) en absence ou en présence d'inhibiteurs spécifiques. *C.R. Acad. Sci. Paris D* **274**, 715–718 (1972).
- Kohler, M. & Schwartz, W. Untersuchungen über die Symbiose von Tieren mit Pilzen und Bakterien. *Z. Allg. Mikro.* **2**, 190–208 (1962).
- Louis, C. & Giannotti, J. Formation lysosomales et membranaires associées des symbiotes et des mycoplasmes endocellulaires d'insectes. *Bull. Obs. Mer.* **43**–55 (1974).
- Du, Y., Maslov, D. A. & Chang, K. Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. *Proc. Natl Acad. Sci. USA* **91**, 8437–8441 (1994).
- Douglas, A. E. Mycetocyte symbiosis in insects. *Biol. Rev.* **64**, 409–434 (1989).
- Smith, D. C. & Douglas, A. E. *The Biology of Symbiosis* (eds Willis, A. J. & Sleight, M. A.) (Edward Arnold, London, 1987).
- Moran, N. & Baumann, P. Phylogenetics of cytoplasmically inherited microorganisms of arthropods. *Trends Ecol. Evol.* **9**, 15–20 (1994).
- Andersson, S. G. E. & Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263–278 (1998).

20. Moran, N. A. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
21. Wernegreen, J. J. & Moran, N. A. Evidence for genetic drift in endosymbionts (*Buchnera*): Analyses of protein-coding genes. *Mol. Biol. Evol.* **16**, 83–97 (1999).
22. Spaulding, A. W. & von Dohlen, C. D. Psyllid endosymbionts exhibit patterns of co-speciation with hosts and destabilizing substitutions in ribosomal RNA. *Insect Mol. Biol.* **10**, 57–67 (2001).
23. Lambert, J. D. & Moran, N. A. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **95**, 4458–4462 (1998).
24. Clark, M. A., Moran, N. A. & Baumann, P. Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol. Biol. Evol.* **16**, 1586–1598 (1999).
25. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86 (2000).
26. Komaki, K. & Ishikawa, H. Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. *J. Mol. Evol.* **48**, 717–722 (1999).
27. Van Ham, R. C. et al. Postsymbiotic plasmid acquisition and evolution of the repA1-replicon in *Buchnera aphidicola*. *Proc. Natl Acad. Sci. USA* **97**, 10855–10860 (2000).
28. Sunnucks, P. & Hales, D. F. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Strobion* (Hemiptera: Aphididae). *Mol. Biol. Evol.* **13**, 510–524 (1996).
29. Ausubel, F. et al. eds *Current Protocols in Molecular Biology* (Wiley, New York, 1996).
30. Grünwald-Janhö, S., Keesey, J., Leous, M., van Miltenburg, R. & Schroeder, C. (eds) *Nonradioactive In situ Hybridization Application Manual* (Boehringer Mannheim, Biochemica, Mannheim, 1996).

Acknowledgements

We thank D. Miller for mealybug identification, D. DeWald, H. Leary, K. Patch, G. Podgorski, J. Shope, A. Spaulding, J. Takemoto and N. Youssef for advice and/or technical help, and R. Andrus, C. Brammer, D. DeWald, K. Patch, E. Pilgrim, G. Podgorski, C. Rowe, L. Smith, and P. Wolf for reviewing the manuscript. This work was supported by a grant from the United States Department of Agriculture and by the Utah Agricultural Experiment Station.

Correspondence and requests for materials should be addressed to C.D.v.D. (e-mail: cvond@biology.usu.edu).

The buffer effect and large-scale population regulation in migratory birds

Jennifer A. Gill*, Ken Norris†, Peter M. Potts‡, Tómas Grétar Gunnarsson§, Philip W. Atkinson¶ & William J. Sutherland||

* Tyndall Centre, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

† School of Animal & Microbial Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK

‡ Solent Court Cottage, Chilling Lane, Warsash, Southampton SO3 9HF, UK

§ Institute of Biology, University of Iceland, Grensásvegur 12, IS 108 Reykjavík, Iceland

|| Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

¶ British Trust for Ornithology, The Nunnery, Thetford IP24 2PU, UK

Buffer effects occur when sites vary in quality and fluctuations in population size are mirrored by large changes in animal numbers in poor-quality sites but only small changes in good-quality sites. Hence, the poor sites 'buffer' the good sites^{1,2}, a mechanism that can potentially drive population regulation if there are demographic costs of inhabiting poor sites. Here we show that for a migratory bird this process can apply on a country-wide scale with consequences for both survival and timing of arrival on the breeding grounds (an indicator of reproductive success^{3,4}). The Icelandic population of the black-tailed godwit, *Limosa limosa islandica*, wintering in Britain has increased fourfold since the 1970s (ref. 5) but rates of change within individual estuaries have varied from zero to sixfold increases. In accordance with the buffer effect, rates of increase are greater on estuaries with low initial numbers, and godwits on these sites have lower prey-intake rates, lower survival rates and arrive later in Iceland than godwits

on sites with stable populations. The buffer effect can therefore be a major process influencing large-scale population regulation of migratory species.

The buffer effect is ecologically important because it describes the mechanism underlying density-dependent changes in fecundity and mortality, and hence provides an insight into processes regulating population growth and abundance. As population density increases, an increasing proportion of animals is displaced into poorer-quality sites leading to reduced fecundity or survivorship. The buffer effect has been demonstrated in a wide range of taxa^{1,6–9}, illustrating a widespread link between increasing population density, distribution patterns across a habitat-quality gradient and demography in animals. Furthermore, there has been considerable theoretical work on distribution patterns and how these are shaped by habitat quality and competition between animals based on the ideal free distribution¹⁰. Despite all this work, it remains unclear whether the buffer effect operates at large spatial scales. This is partly because information is lacking, but also because it is possible that it becomes more difficult for animals to track the quality of a number of different sites, or to be able to travel without great cost between sites, as spatial scale increases. Also, animal populations that exist over large spatial scales are often migratory. This raises the possibility that any buffer effect might impact on demography in several stages of the annual cycle, and in geographical areas outside the range within which it is evident.

To examine these ideas, we studied during the non-breeding period a population of migratory shorebirds that has undergone a significant increase in population size over the past 30 years. The Icelandic-breeding population of black-tailed godwits spends the non-breeding period feeding on estuaries principally in Ireland, the UK and France. To critically examine whether this population shows evidence of a buffer effect, we addressed several specific questions. First, as population size has increased over time, has the increase been greatest on sites that initially had only small populations of

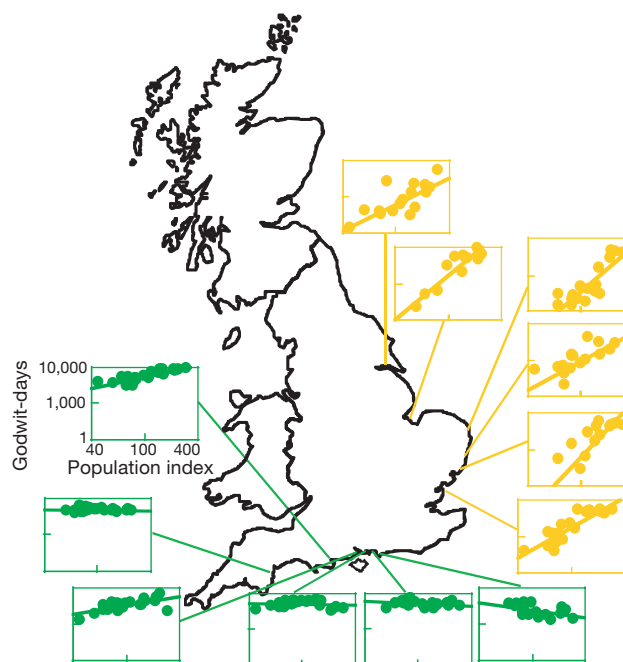


Figure 1 Rates of change of godwit numbers differ at different sites. Numbers of black-tailed godwits on individual wintering sites on the south and east coasts are shown relative to the godwit population index for Britain and Northern Ireland between 1970 and 1996 (data from the Wetland Bird Survey⁵). Axes are identical on all graphs. Number of godwit-days = monthly count × number of days each month, summed from October to March.

birds? At low population sizes the buffer effect predicts that most birds occupy the best-quality sites. As population size increases, numbers increase most rapidly on poor-quality sites. This results in a negative correlation between initial population size on a site, and the rate of increase in numbers on it as overall population size increases. Second, are any such changes being driven by the overall population increase or has the quality of different sites changed over time? The buffer effect assumes that differences in quality between sites are maintained over time. Third, do birds occupying sites that experience the greatest increase in numbers as overall population size increases have lower rates of food intake, poorer survival and later arrival dates in Iceland? The buffer effect predicts that such sites should be of poorer quality and that this should influence demography.

The long-term monitoring of waterfowl through the Wetland Bird Survey⁵ is based on monthly counts carried out at most coastal and many wetland sites in the UK. To allow for missing counts in any given month, an index of population size that imputes missing data (the Underhill index¹¹) is widely used to allow between-year comparisons of population size. Within the UK, individual wintering sites have undergone very different patterns of population change as the Icelandic population has increased (Fig. 1). The slopes of these relationships give the rate of population increase on each site relative to the national population increase. Under Brown's concept of the buffer effect², poor sites should be avoided at low population densities and population increases will result in expansion into these sites^{12,13}. This process results in a negative relationship between initial population size and subsequent rates of increase. Across all major wintering sites in the UK, there is a strong negative relationship between initial population size on each site in the early 1970s and the rates of increase since then (Fig. 2). The general, but not universal, trend is for populations on estuaries on the east coast of England to have increased rapidly, whereas those on the other major wintering area in England, the south coast, have remained largely stable. Although population densities on these sites cannot be calculated because the area of suitable habitat is unknown, intertidal area is not related to either initial ($r^2 = 0.0001$, NS (not significant)) or current ($r^2 = 0.001$, NS) population sizes on each estuary, indicating that small populations do not simply occur on small estuaries.

For wintering waders, prey-intake rates are critically important

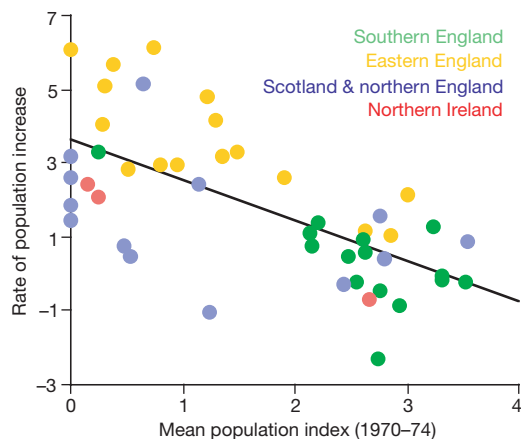


Figure 2 Godwit populations that are initially smaller grow more rapidly. The rate of change in numbers of godwit-days on individual sites (that is, the slopes of the relationships in Fig. 1) is plotted against the mean population index on those sites between 1970 and 1974 ($y = 3.66 - 1.1x$, $r^2 = 0.43$, $P < 0.0001$). Within each of the four regions presented each relationship is also negative, although only two trends are statistically significant (Spearman rank correlations: southern England, $r_s = -0.56$, $n = 15$, $P < 0.04$; eastern England, $r_s = -0.73$, $n = 16$, $P < 0.005$; Scotland and northern England, $r_s = -0.49$, $n = 13$, NS; Northern Ireland, $r_s = -1$, $n = 3$, NS).

for survival and migration and are therefore a major factor determining habitat choice¹⁴. We determined the seasonal changes in prey-intake rates of black-tailed godwits at four south and five east coast wintering sites. At the start of winter on all these sites, godwits forage on estuarine mudflats at which prey depletion results in a seasonal decline in prey-intake rates¹⁵. At some sites (particularly on the south coast), the birds are able to respond to prey depletion by switching from poor-quality estuarine sites to exploit increasing prey densities, principally earthworms (Lumbricidae) on recently flooded freshwater meadows. The timing of the flooding of freshwater sites in relation to prey depletion on the estuaries is likely to be the factor that constrains numbers of godwits on south coast sites. This switch inland from poor sites results in large differences between east and south coast sites in prey-intake rates at the end of the winter, with prey-intake rates being significantly higher at south coast sites (mean $\pm$ s.e.m. = 0.46 ± 0.06 mg ash-free dry mass (AFDM) s^{-1}) than at east coast sites (0.17 ± 0.04 mg AFDM s^{-1} , Mann-Whitney $U_{4,5} = 20$, $P < 0.015$). Spring prey-intake rates are strongly negatively related to the rates of population change on those sites (Fig. 3a). Thus, sites with stable populations have higher spring prey-intake rates than sites at which godwit numbers are rapidly increasing. Differences in prey-intake rates between the east and south coasts may therefore be the mechanism driving differences in population trajectories. An alternative explanation for the pattern of increasing populations on east coast sites could be that the quality of east coast sites has changed through time. Invertebrate

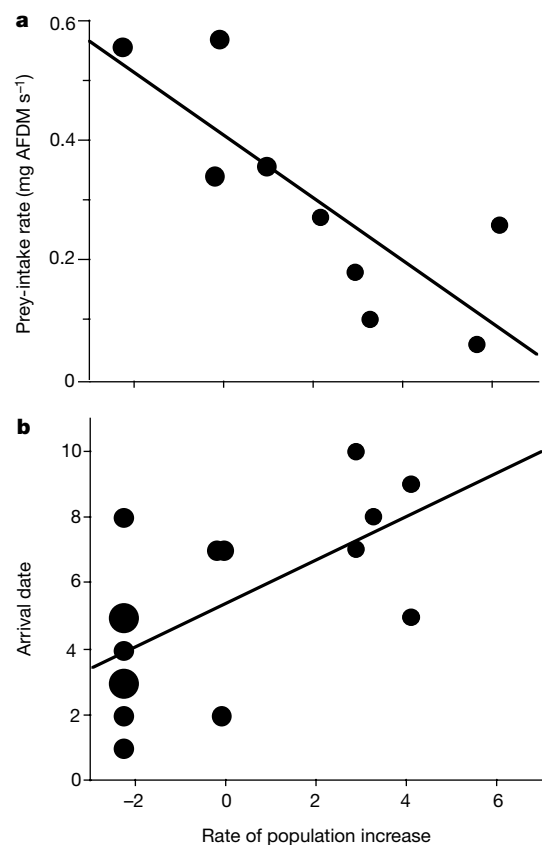


Figure 3 Godwits from increasing populations have lower prey-intake rates and arrive later in Iceland. **a**, The rate of population change on nine individual wintering sites plotted against mean prey-intake rate in March ($y = 5.82 - 12.53x$, $r^2 = 0.66$, $P < 0.009$; for one south coast estuary, the intake rate is from the freshwater meadows as all godwits from this estuary switch to the meadows in January). **b**, The rate of population change plotted against the arrival dates (days since 20 April) in Iceland of individually marked birds from those populations ($y = 5.43 + 0.65x$, $r^2 = 0.66$, $P < 0.01$). Larger dots represent two individuals.

abundance was measured with the same methodology at 83 sampling locations across four of the east coast estuaries during both the 1970s (ref. 16) and this study in the 1990s. The abundance of the bivalve species consumed by the godwits did not differ significantly between the two periods ($F_{3,158} = 1.30$, NS).

Are higher prey-intake rates related to increased survival? On the south and east coasts of England, black-tailed godwits have been individually colour-ringed for the past 8 and 6 years, respectively. Survival models for adult godwits over the past 6 years (1995–2000) were fitted by the Cormack–Jolly–Seber method in the computer package MARK¹⁷. Both survival (ϕ) and reporting rates (λ) varied with time (t , annually) and by group (g ; south or east coast wintering birds; data from individual sites were not sufficient to carry out a site-based survival analysis). We ran 16 models, from the most complex, $\phi_{gt}\lambda_{gt}$, to the simplest, $\phi\lambda$, and all intermediates (Table 1). Model fit was assessed with the Akaike Information Criterion (AIC), the lowest value of which indicates the model that describes the data most parsimoniously. For nested models a likelihood ratio test was used to confirm the results obtained with the AIC and the simpler model was rejected if a significant result was obtained.

As is common with such frequency data, the pooled chi-squared goodness-of-fit test indicated that the data were over-dispersed in respect of the resighting data. Hence, the model selection data were modified so that the corrected quasi-likelihood AIC (QAICc) was used to select the most parsimonious model. This adjustment incorporated a measure of the over-dispersion ($\hat{c}$) calculated from the goodness-of-fit test by dividing the observed model by the mean pooled deviance, resulting in $\hat{c} = 1.51$. The most parsimonious model was $\phi_g\lambda_{gt}$ (Table 1), with group-dependent survival and both time- and group-dependent reporting parameters. Average annual survival was significantly higher for birds wintering on the south coast (0.94 ± 0.02 s.e.m.) than for those on the east coast (0.87 ± 0.02 s.e.m.; comparison of models 1 and 2 in Table 1, $\chi^2_1 = 4.34$, $P < 0.05$). Adding time dependence into the models did not further improve the model fit (Table 1; comparison of models 1 and 4, $\chi^2_8 = 11.36$, $P = 0.18$).

Prey-intake rates at wintering locations could affect subsequent breeding success by influencing the timing of arrival on the breeding grounds (through processes such as optimal territory acquisition¹⁸ and seasonal declines in clutch size¹⁹ and offspring survival²⁰) and/or the condition of the birds on arrival. We examined arrival dates of individually marked godwits during late April and early May 2000. On arrival in Iceland, godwits congregate in large flocks on a small number of estuarine and wetland sites. Regularly monitoring these arrival sites, we located 60 individually marked godwits; for 16 of these individuals, the site they had inhabited during the winter of 1999–2000 was known. For these birds it was therefore possible to

assess the relationship between the rates of population increase on their wintering sites and the individual arrival dates. Birds from sites with stable populations (and high spring prey-intake rates) arrived in Iceland significantly earlier than those from sites with rapidly increasing populations (and lower spring prey-intake rates; Fig. 3b). The quality of wintering sites may therefore not only influence survival and the condition of birds locally, but may also influence breeding success in locations $>2,000$ km distant. Although the cause of the increase in the godwit population is unknown, the expansion into poorer-quality wintering sites shows that it is clearly not the result of improved winter feeding conditions.

The expanding godwit population exhibits a clear large-scale buffer effect, in which numbers in poor sites have undergone disproportionately high rates of increase, and site quality is related to both survival and timing of arrival on the breeding grounds. The disproportionate increase in numbers in poor locations therefore has the potential to influence patterns of population regulation. The buffer effect is a common phenomenon in a wide range of taxa but its capacity to regulate populations over large spatial scales has never been addressed. This study has shown that buffer effects operating in one season can influence fitness parameters at different stages of the annual life cycle and across breeding and wintering ranges. For migratory species, buffer effects may therefore be a major density-dependent process influencing large-scale population regulation. □

Received 27 March; accepted 1 June 2001.

1. Kluver, H. N. & Tinbergen, L. Territory and the regulation of density in titmice. *Arch. Néerl. Zool.* **10**, 265–289 (1953).
2. Brown, J. L. The buffer effect and productivity in tit populations. *Am. Nat.* **103**, 347–354 (1969).
3. Kokko, H. Competition for early arrival in migratory birds. *J. Anim. Ecol.* **68**, 940–950 (1999).
4. Møller, A. P. Phenotype-dependent arrival time and its consequences in a migratory bird. *Behav. Ecol. Sociobiol.* **35**, 115–122 (1994).
5. Pollitt, M. S. et al. *The Wetland Bird Survey 1998–99: Wildfowl and Wader Counts* (BTO, WWT, RSPB & JNCC, Slimbridge, 2000).
6. Krebs, J. R. Regulation of numbers in the Great Tit (*Aves: Passeriformes*). *J. Zool.* **162**, 317–333 (1970).
7. Ferrer, M. & Donazar, J. A. Density-dependent fecundity by habitat heterogeneity in an increasing population of Spanish Imperial Eagles. *Ecology* **77**, 69–74 (1996).
8. Fraser, D. & Sise, T. Observations on stream minnows in a patchy environment: a test of theory of habitat selection. *Ecology* **61**, 790–797 (1980).
9. Halama, K. J. & Dueser, R. D. Of mice and habitats: tests for density-dependent habitat selection. *Oikos* **69**, 107–114 (1994).
10. Fretwell, S. D. & Lucas, H. J. Jr. On territorial behaviour and other factors influencing habitat distribution in birds. *Acta Biotheor.* **19**, 16–36 (1970).
11. Underhill, L. G. & Prys-Jones, R. Index numbers for waterbird populations. I. Review and methodology. *J. Appl. Ecol.* **31**, 463–480 (1994).
12. Sutherland, W. J. *From Individual Behaviour to Population Ecology* (Oxford Univ. Press, Oxford, 1996).
13. Moser, M. E. Limits to the numbers of grey plovers *Pluvialis squatarola* wintering on British estuaries: an analysis of long-term population trends. *J. Appl. Ecol.* **25**, 473–485 (1988).
14. Goss-Custard, J. D. Competition for food and interference among waders. *Ardea* **68**, 31–52 (1980).
15. Gill, J. A., Sutherland, W. J. & Norris, K. Depletion models can predict shorebird distribution at different spatial scales. *Proc. R. Soc. Lond. B* **268**, 369–376 (2001).
16. Kay, D. G. & Knights, R. D. The macro-invertebrate fauna of the intertidal soft sediments of south east England. *J. Mar. Biol. Assoc. UK* **55**, 811–832 (1975).
17. White, G. C. & Burnham, K. P. Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**, S120–S139 (2000).
18. Currie, D., Thompson, D. B. A. & Burke, T. Patterns of territory settlement and consequences for breeding success in the Northern Wheatear *Oenanthe oenanthe*. *Ibis* **142**, 389–398 (2000).
19. Sandercock, B. K., Lank, D. B. & Cooke, F. Seasonal declines in the fecundity of arctic-breeding sandpipers: different tactics in two species with an invariant clutch size. *J. Avian Biol.* **30**, 460–468 (1999).
20. Lepage, D., Gauthier, G. & Menu, S. Reproductive consequences of egg-laying decisions in snow geese. *J. Anim. Ecol.* **69**, 414–427 (2000).

Acknowledgements

We are grateful for the financial and technical support of NERC; the TMR programme of the EC through Sandgerdi Marine Institute, Iceland; and A. Gardarsson and the Icelandic Institute of Natural History. We are grateful to L. Norton, J. Smart, D. McNeill, R. Croger, G. Appleton and T. Turner for assistance in the field, and to I. Côté, J. Reynolds and A. Watkinson for comments on the manuscript. Survival data were generated from Farlington Ringing Group and Wash Wader Ringing Group colour-ringing schemes. The FRG scheme is partly funded by the Farlington Ringing Group, Hampshire Ornithological Society and Hampshire County Council. Wetland Bird Survey data were provided by the British Trust for Ornithology.

Correspondence and requests for materials should be addressed to J.A.G. (e-mail: j.gill@uea.ac.uk).

Table 1 Survival models of adult black-tailed godwits wintering on the east and south coasts of England

No.	Model	QAICc	No. of parameters	Deviance
1	$\phi_g\lambda_{gt}$	954.23	12	118.50
2	$\phi\lambda_{gt}$	955.05	11	122.84
3	$\phi_t\lambda_{gt}$	959.11	15	116.52
4	$\phi_{gt}\lambda_{gt}$	963.29	20	107.14
5	$\phi_t\lambda_g$	964.17	7	148.94
6	$\phi_{gt}\lambda_g$	964.24	12	133.61
7	$\phi_g\lambda_t$	978.99	7	171.32
8	$\phi_{gt}\lambda$	988.18	11	172.86
9	$\phi_{gt}\lambda_t$	991.37	15	165.23
10	$\phi\lambda_t$	995.37	6	199.12
11	$\phi_t\lambda$	997.51	6	202.34
12	$\phi_t\lambda_t$	1000.53	10	194.60
13	$\phi\lambda_g$	1057.25	3	301.71
14	$\phi_g\lambda_g$	1058.91	4	301.17
15	$\phi_g\lambda$	1072.00	3	323.98
16	$\phi\lambda$	1084.45	2	345.82

All 16 possible models are presented and ordered by the quasi-likelihood Akaike Information Criterion (QAICc).

Crossflow filtration in suspension-feeding fishes

S. Laurie Sanderson*, Angela Y. Cheer†, Jennifer S. Goodrich*‡, Jenny D. Graziano*‡ & W. Todd Callan*

* Department of Biology, College of William and Mary, Williamsburg, Virginia 23187-8795, USA

† Department of Mathematics and Institute of Theoretical Dynamics, University of California, Davis, California 95616-8618, USA

Rows of comb-like or tufted gill rakers in the oral cavity of suspension-feeding fishes (for example, herring, anchovies and tilapia) have been thought to serve as (1) non-porous barriers that direct particle-laden water to the sticky oral roof, where particles are retained as water exits from the oral cavity, (2) conventional dead-end filters that sieve particles from water exiting between rakers, or (3) sticky filters that retain particles encountered by a hydrosol filtration mechanism^{1–6}. Here we present data from computational fluid dynamics and video endoscopy in suspension-feeding fish indicating that the rakers of three distantly related species function instead as a crossflow filter^{7,8}. Particles are concentrated inside the oral cavity as filtrate exits between the rakers, but particles are not retained on the rakers. Instead, the high-velocity crossflow along the rakers carries particles away from the raker surfaces and transports the particles towards the oesophagus. This crossflow prevents particles from clogging the gaps between the rakers, and solves the mystery of particle transport from the rakers to the oesophagus.

Our numerical simulations showed that the rakers of pump suspension-feeding fishes operating at moderate Reynolds numbers ($Re = 150–600$) are always leaky, never serving as a non-porous barrier to flow (see Supplementary Information). Leakiness (defined as the ratio of the volume of fluid that actually passes through the unblocked area between a pair of rakers per unit time to the volume that would pass through an equivalent area per unit time in the absence of the rakers^{9,10}) ranged from 0.45 to 0.79 (Table 1), rejecting the hypothesis^{6,11} that rakers function as a non-porous barrier to fluid flow at these Reynolds numbers. This result was obtained for biologically relevant raker shapes (conical ellipses and elliptical cylinders), raker sizes (250–500 μm minor axis, 750–1500 μm major axis), gap widths between the tips of rakers (0–750 μm), velocities at which flow approached the rakers (35–70 cm s^{-1}), angles of approaching flow (75° from the positive z -axis and 60° from the positive x -axis), and a finite as well as an infinite row of rakers (Fig. 1a–c). Flow between the rakers also occurred in the presence of non-porous blockers spaced at intervals to simulate the protuberances found between rakers in some species (Fig. 1d, e).

As the rakers are leaky, they could theoretically function as dead-end sieves, with the gaps between rakers serving as the pores of the sieve. As in a colander or tea strainer, particles that are too large to pass through the pores of a dead-end sieve are retained on the filter surface when water exits perpendicular to the sieve. As a supplement or alternative to sieving, particles smaller than the gaps between rakers could be retained by entrapment in sticky mucus on the rakers by a hydrosol filtration mechanism^{12,13}.

Our experiments with a miniature fibre-optic endoscope inserted into the oral cavity of gizzard shad (*Dorosoma cepedianum*, Clupeidae), goldfish (*Carassius auratus*, Cyprinidae) and ngege tilapia (*Oreochromis esculentus*, Cichlidae) tested the hypotheses

Table 1 Leakiness calculated from numerical simulations

	Infinite row of rakers		Row of three rakers
	Elliptical cylinders*	Conical ellipses†	Conical ellipses‡
Unblocked	0.79 (195)‡	0.68 (145)	0.70 (150)
Blocked	0.53 (255)	0.45 (185)	0.45 (175)

The total magnitude of the flow velocity entering the grid was 35 cm s^{-1} .

* 500 μm minor axis, 1,500 μm major axis, 750 μm gap between rakers.

† 500 μm minor axis, 1,500 μm major axis, 100 μm gap between raker tips.

‡ All Reynolds numbers (in parentheses) were calculated from the downstream length (minor axis) of the rakers and the maximum total magnitude of the filtrate flow between rakers.

that rakers function as dead-end sieves or as sticky filters. In endoscopic videotapes recorded during suspension feeding by these unrestrained fish, 95% of the particles travelled posteriorly towards the oesophagus without contacting any oropharyngeal surface (Fig. 2). High-speed video endoscopy (50–250 frames s^{-1}) with an intensified imager in the ngege tilapia permitted tracking of the predominant particle path towards the posterior oral cavity in slow motion, confirming previous observations made at 30 frames s^{-1} (ref. 14). In all three species, only 1% of the particles rested on oral surfaces for a mean of 0.1 s (range 0.07–0.47 s, $n = 9$ particles) before lifting and continuing posteriorly. As particles rarely contacted the rakers and did not accumulate there, dead-end sieving did not occur.

In contrast to the multiple particles retained in mucous aggregates that moved as a unit in the oral cavity of Nile tilapia⁵, particles in the oral cavities of gizzard shad, goldfish and ngege tilapia moved independently of each other. Whereas particles retained in mucus were prevalent on the rakers of Nile tilapia⁵ and on the oral roof of blackfish⁶, mucus entrapment did not occur in gizzard shad, goldfish or ngege tilapia (Fig. 2).

Together, the numerical simulations and endoscopic videos indicate that the rakers of these three species and of blackfish function as a crossflow filter (Fig. 3). Crossflow filtration is distinct

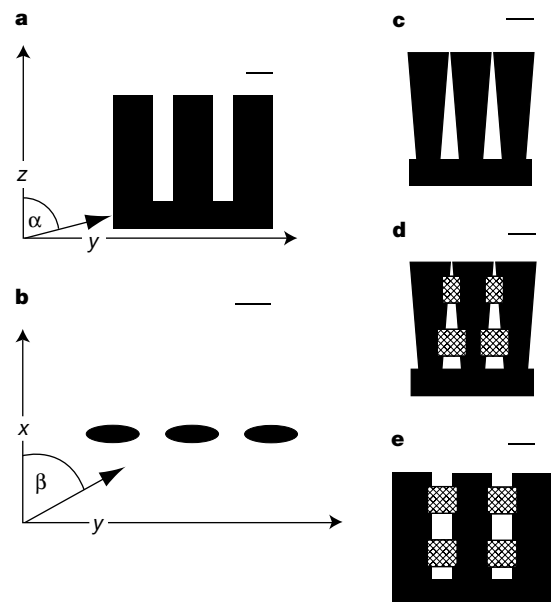


Figure 1 Examples of raker morphologies used in the numerical simulations. **a**, Lateral view of elliptical cylinders, with the angle of approaching flow indicated by $\alpha = 75^\circ$ from the positive z -axis. **b**, Transverse section of elliptical cylinders, with the angle of approaching flow indicated by $\beta = 60^\circ$ from the positive x -axis. **c**, Conical ellipses, lateral view. **d**, Conical ellipses with non-porous blockers, lateral view. **e**, Elliptical cylinders with non-porous blockers, lateral view. Scale bars, 500 μm .

‡ Present addresses: Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA (J.S.G.); College of Medicine, Howard University, Washington, DC 20059, USA (J.D.G.).

from dead-end sieving in that the fluid being filtered is pumped along the surface of the filter (for example, parallel to the surface) rather than perpendicular to the surface of the filter. Whereas filtrate turns to exit through the pores of the filter, the crossflow along the filter tends to remove particles from the immediate vicinity of the filter surface and transport the concentrated particles downstream. Within the past 30 years, crossflow filtration has become prevalent in a wide range of industries (for example, water purification, food and beverages, and biotechnology) because particles can be separated economically with minimal accumulation on the filter⁸. In pump suspension-feeding fishes, a high-velocity crossflow ($\sim 55 \text{ cm s}^{-1}$) sweeping across the raker surfaces⁶ transports particles towards the oesophagus (gizzard shad, goldfish and ngege tilapia) or towards the oral roof (blackfish). Particles are carried posteriorly with this crossflow along the raker surfaces, rather than turning with the flow of filtrate that exits between the rakers (Figs 2 and 3). Thus, as the oral cavity of the fish narrows posteriorly, particles are concentrated but remain suspended in the crossflow.

Crossflow filtration in suspension-feeding fishes differs from most industrial crossflow filtration in that, within the particle size range studied ($40 \mu\text{m}$ to 1 mm), particles do not accumulate on the rakers. During most industrial crossflow filtration, the flow of filtrate through the filter (termed filtrate flux) is restricted eventually by the deposition of a thin layer of macromolecules, colloids, or fine or large particles on the filter^{15–17}. However, the filtrate flux for industrial suspensions of particles larger than $\sim 1 \text{ nm}$ is higher than that predicted by the gel-polarization model, in which brownian diffusion alone causes back-transport of particles from the filtration surface into the crossflow^{18,19}. This 'flux paradox' for suspensions of colloids and larger particles has been partially resolved by models incorporating additional mechanisms that increase back-transport of particles and/or filtrate flux, such as inertial lift, shear-induced diffusion, concentrated flowing layers and surface transport^{20–22}.

Our calculations of radial inertial migration²³ of particles (40--

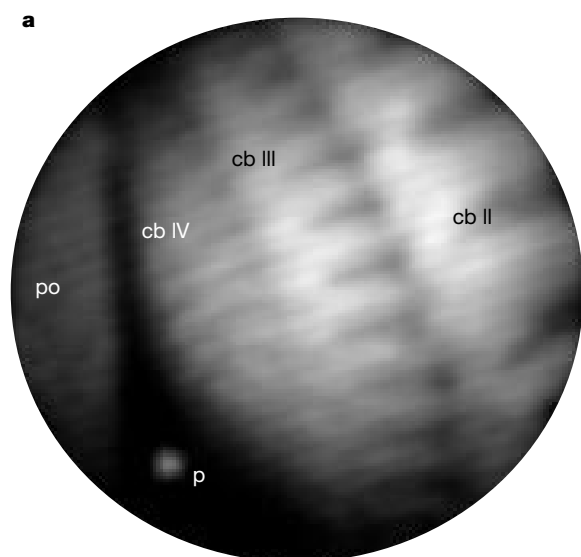


Figure 2 The gill rakers of goldfish, gizzard shad and ngege tilapia do not function as a dead-end sieve or as a sticky filter. **a**, Endoscopic image of goldfish oral cavity during suspension feeding, showing the left ceratobranchial arches IV (cb IV), III (cb III) and II (cb II) to which the rakers are attached, the fleshy palatal organ (po) on the oral roof, and a particle (p, $\sim 250 \mu\text{m}$ diameter). The posterior oral cavity is towards the top of the image. **b**, Movement frequencies of 100 randomly selected particles ($38 \mu\text{m}$ to 1 mm diameter) in the oral cavities of each of three specimens belonging to each of the three species.

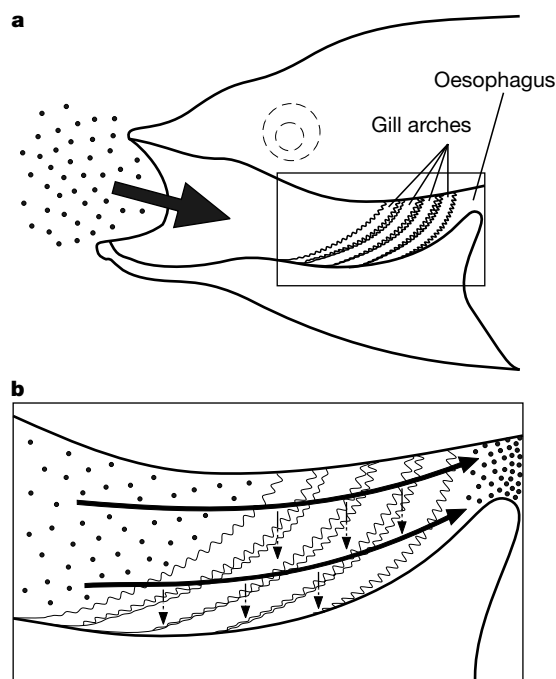
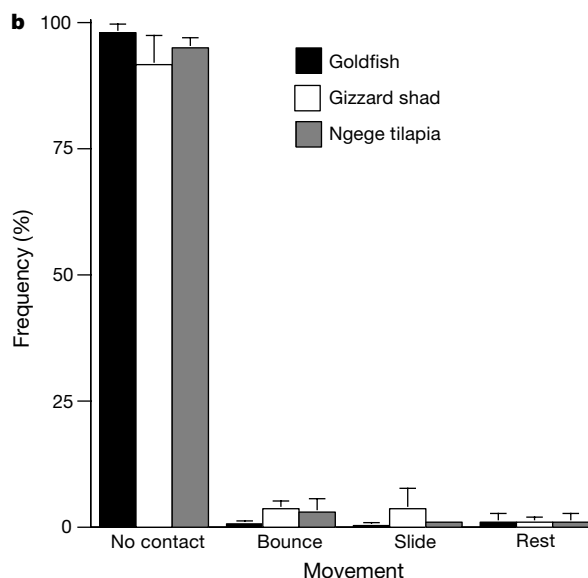


Figure 3 During crossflow filtration, the fluid to be filtered is pumped along the surface of the filter as filtrate exits through the filter pores. **a**, Sagittal view of the oral cavity, illustrating the flow (arrow) of water and particles entering the oral cavity. The rakers, which serve as the filter, are attached to the gill arches. **b**, Enlargement of gill-arch area showing particles being concentrated near the oesophagus as they are carried with the crossflow (large arrows) towards the posterior oral cavity rather than travelling with the flow of filtrate (small arrows), which exits from the oral cavity by passing between the rakers. Owing to lower pressure outside the oral cavity relative to inside at this stage of feeding^{24,25}, the filtrate diverges from the crossflow. For clarity, particles have not been drawn travelling over the arches.



Movement was classified as (1) no contact, particle moved from the anterior of the oral cavity to the posterior without touching any oropharyngeal surface; (2) bounce, particle bounced off the floor or roof of the oral cavity and proceeded posteriorly; (3) slide, particle maintained contact with the arches and rakers as it moved posteriorly; or (4) rest, particle remained immobile on an arch, a raker or the oral roof for two or more frames before proceeding posteriorly. Values are mean $\pm$ s.d.

100 μm diameter) away from the rakers in a 'tubular pinch effect' indicate that inertial lift is at least one order of magnitude too low to account for the lack of contact with the rakers. (From approximations based on limited data available in the literature^{6,24,25}, channel $\text{Re} = 1,750$, wall shear rate = $1,000 \text{ s}^{-1}$ and transraker pressure = 10 kPa.) No endoscopic evidence exists for shear-induced diffusion or concentrated flowing layers in suspension-feeding fishes, and only limited surface transport occurs (Fig. 2b). Further study will assess the importance of additional processes that could operate in conjunction with inertial lift to limit particle deposition on the rakers during crossflow filtration in suspension-feeding fishes, including pulsatile flow, flow reversals, vortices and fluid mixing due to surface roughness^{7,26}.

The role of inter-raker spacing in determining the size of prey retained by planktivorous fishes is controversial^{27–29}. As the species that we studied do not use dead-end sieving to retain particles, the gap widths between rakers do not necessarily serve as prey-size thresholds. We propose that the spaces between the rakers are as large as possible while still allowing prey to be diverted from the filtrate streamlines to the crossflow streamlines. This perspective shifts the focus from particle size as a mechanical threshold for retention to particle size as a hydrodynamic threshold that affects the magnitude of the lift and shear^{20–22} experienced by particles at the boundary between the crossflow and the filtrate flow. Although the flow velocity of the filtrate between the rakers during feeding cannot be measured with current technology, we are testing this model using numerical simulations, physical models and *in vivo* experiments with particles that vary in size or density. □

Methods

Numerical simulations

A flow solver (FLOW-3D, Flow Science) was used to simulate three-dimensional, incompressible, viscous flow. The numerical method was a finite-difference approximation to the Navier–Stokes equations of motion, with pressure and velocity coupled implicitly. The resulting set of algebraic equations was solved by successive over-relaxation or alternating-direction implicit methods.

The simulated rakers were oriented so that (1) flow in the x -direction passed between the rakers to exit the oral cavity; (2) flow in the y -direction travelled parallel to the major axis of the rakers, heading towards the posterior oral cavity; and (3) flow in the z -direction travelled along the height of the rakers. For the simulations in which the total magnitude of the flow velocity entering the grid was 35 cm s^{-1} , the x -component of the flow velocity entering the grid was 17.047 cm s^{-1} , the y -component was 29.526 cm s^{-1} , and the z -component was 7.911 cm s^{-1} . The low x - and low y -boundaries were fixed-velocity inflow boundaries and the high x - and high y -boundaries were free-exit (continuation) boundaries. A symmetry boundary was placed at $z = 0$ and at the highest z -value. To model an infinite row of rakers, the gill arch to which the rakers were attached was extended by one-half of the width of the gap between the rakers, and periodic boundaries were set at the y -values at the ends of the gill arch. The number of mesh cells used for the infinite row of rakers was 23 in the x -direction, 11 in the y -direction, and 10 in the z -direction. For the finite row of three rakers, the number of cells in the y -direction was increased to 64.

Video endoscopy

A polyethylene cannula (2.15 mm inner diameter, 3.25 mm outer diameter, Intramedic PE 280) was implanted into the oral cavity^{5,6,14} through a hole in the left preopercular bone of three gizzard shad (28.2–29.2 cm standard length), three goldfish (15.2–16.5 cm standard length) and three ngege tilapia (22.3–25.8 cm standard length). About 4 h after each fish had been returned to its 130-l aquarium, a flexible fibre-optic endoscope (Olympus ultrathin fiberscope type 14, 1.4 mm outer diameter, 1.2 m working length, 75° field of view, 0.2–5.0 cm depth of field) was threaded through the cannula. The endoscope was attached to a CCD (charge-coupled device) video camera (Canon Ci-20R, 30 frames s^{-1}) or a Kodak Ektapro Hi-Spec Motion Analyzer 1012/2 with an Intensified Imager VSG (50–250 frames s^{-1}) connected to a Hi-8 video player/recorder with a jog shuttle (Sony EVO-9700), which permitted frame-by-frame analysis of videotapes. A high-intensity light source (Olympus Heliod ALS-6250, 250 W) provided light to the endoscope.

About 80% of the oropharyngeal cavity was visible through the endoscope, including the ceratobranchials and associated rakers of arches I–IV. In the goldfish and the ngege tilapia, the fleshy oral roof could also be seen. In the gizzard shad, which lacks a fleshy oral roof, the epibranchials and associated rakers of arches I–IV were visible dorsally. To prepare Fig. 2a, Hi-8 videotape was digitized with a RasterOps 24 STV colour display board. The image was then processed by convolving with a mean kernel (5 × 5 pixels) using

NIH Image 1.52, to remove the fine honeycomb pattern caused by individual fibres in the endoscope's fibre-optic bundle.

Videotapes were recorded while the fish fed on a dilute suspension (volume concentration ~0.1–1%) of finely crushed Tetramin flakes (0.1–1.0 mm diameter), brine shrimp cysts (*Artemia* sp., 210–300 μm diameter) and polystyrene microspheres (BioRad Bio-Beads SM-2, 38–75 μm diameter). These particles could be discerned clearly in endoscopic videotapes during feeding^{5,6,14}. Experiments were conducted at 23.0–29.5 °C. At the conclusion of each experiment, the cannula was removed; the insertion site healed subsequently.

Received 5 February; accepted 29 May 2001.

- Kardong, K. V. *Vertebrates: Comparative Anatomy, Function, Evolution* 2nd edn (McGraw-Hill, Boston, 1998).
- Gerking, S. D. *Feeding Ecology of Fish* (Academic, San Diego, 1994).
- Vogel, S. *Life in Moving Fluids: The Physical Biology of Flow* 2nd edn (Princeton Univ. Press, Princeton, 1994).
- Hoogenboezem, W. et al. A new model of particle retention and branchial sieve adjustment in filter-feeding bream (*Abramis brama*, Cyprinidae). *Can. J. Fish. Aquat. Sci.* **48**, 7–18 (1991).
- Sanderson, S. L. et al. Mucus entrapment of particles by a suspension-feeding tilapia (Pisces: Cichlidae). *J. Exp. Biol.* **199**, 1743–1756 (1996).
- Sanderson, S. L., Cech, J. J. Jr & Patterson, M. R. Fluid dynamics in suspension-feeding blackfish. *Science* **251**, 1346–1348 (1991).
- Belfort, G., Davis, R. H. & Zydney, A. L. The behavior of suspensions and macromolecular solutions in crossflow microfiltration. *J. Membr. Sci.* **96**, 1–58 (1994).
- Zeman, L. J. & Zydney, A. L. *Microfiltration and Ultrafiltration: Principles and Applications* (Dekker, New York, 1996).
- Cheer, A. Y. L. & Koehl, M. A. R. Paddles and rakes: fluid flow through bristled appendages of small organisms. *J. Theor. Biol.* **129**, 17–39 (1987).
- Koehl, M. A. R. in *Biological Fluid Dynamics, Soc. Exp. Biol. Symp.* Vol. 49 (eds Ellington, C. P. & Pedley, T. J.) 157–182 (Company of Biologists, London, 1995).
- Sanderson, S. L. & Cheer, A. Y. in *Fluid Dynamics in Biology* (eds Cheer, A. Y. & van Dam, C. P.) 135–160 (American Mathematical Society, Providence, 1993).
- Shimeta, J. & Jumars, P. A. Physical mechanisms and rates of particle capture by suspension-feeders. *Oceanogr. Mar. Biol. Annu. Rev.* **29**, 191–257 (1991).
- LaBarbera, M. Feeding currents and particle capture mechanisms in suspension feeding animals. *Am. Zool.* **24**, 71–84 (1984).
- Goodrich, J. S., Sanderson, S. L., Batjakas, I. E. & Kaufman, L. S. Branchial arches of suspension-feeding *Oreochromis esculentus*: sieve or sticky filter? *J. Fish Biol.* **56**, 858–875 (2000).
- Li, H., Fane, A. G., Coster, H. G. L. & Vigneswaran, S. Direct observation of particle deposition on the membrane surface during crossflow microfiltration. *J. Membr. Sci.* **149**, 83–97 (1998).
- Chellam, S. & Wiesner, M. R. Particle back-transport and permeate flux behavior in crossflow membrane filters. *Environ. Sci. Technol.* **31**, 819–824 (1997).
- Wiesner, M. R. Morphology of particle deposits. *J. Environ. Eng.* **125**, 1124–1132 (1999).
- Porter, M. C. Concentration polarization with membrane ultrafiltration. *Ind. Eng. Chem. Prod. Res. Develop.* **11**, 234–248 (1972).
- Green, G. & Belfort, G. Fouling of ultrafiltration membranes: lateral migration and the particle trajectory model. *Desalination* **35**, 129–147 (1980).
- Sethi, S. & Wiesner, M. R. Modeling of transient permeate flux in cross-flow membrane filtration incorporating multiple particle transport mechanisms. *J. Membr. Sci.* **136**, 191–205 (1997).
- Tardieu, E., Grasmick, A., Geaugey, V. & Manem, J. Hydrodynamic control of bioparticle deposition in a MBR applied to wastewater treatment. *J. Membr. Sci.* **147**, 1–12 (1998).
- Bowen, W. R. & Jenner, F. Theoretical descriptions of membrane filtration of colloids and fine particles: an assessment and review. *Adv. Colloid Interface Sci.* **56**, 141–200 (1995).
- Drew, D. A., Schonberg, J. A. & Belfort, G. Lateral inertial migration of a small sphere in fast laminar flow through a membrane duct. *Chem. Eng. Sci.* **46**, 3219–3224 (1991).
- Lauder, G. V. The suction feeding mechanism in sunfishes (*Lepomis*): an experimental analysis. *J. Exp. Biol.* **88**, 49–72 (1980).
- Norton, S. F. & Brainerd, E. L. Convergence in the feeding mechanics of ecomorphologically similar species in the Centrarchidae and Cichlidae. *J. Exp. Biol.* **176**, 11–29 (1993).
- Cheryan, M. *Ultrafiltration and Microfiltration Handbook* (Technomic, Lancaster, 1998).
- Mummert, J. R. & Drenner, R. W. Effect of fish size on the filtering efficiency and selective particle ingestion of a filter-feeding clupeid. *Trans. Am. Fish. Soc.* **115**, 522–528 (1986).
- Langeland, A. & Nøst, T. Gill raker structure and selective predation on zooplankton by particulate feeding fish. *J. Fish Biol.* **47**, 719–732 (1995).
- Gibson, R. N. Development, morphometry and particle retention capability of the gill rakers in the herring, *Clupea harengus* L. *J. Fish Biol.* **32**, 949–962 (1988).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We are grateful to J. Perkins, S. Yu, J. Wagner, S. Leman, J. Yang, C. Chesnutt and B. Rourke for assistance with data collection and analysis; G. Belfort, E. L. Brainerd, M. Patterson and M. Wiesner for comments on the manuscript; H. Austin, P. Geer, J. Goins, T. Mathes, J. Owens, M. Roberts, P. Sadler, Z. Stroud, J. Viehweg and M. A. Vogelbein for assistance in specimen collection and care; B. Bunting and H. Burrell for figure preparation; and K. Wienczek for technical advice. This work was supported by grants from the NSF (S.L.S., A.Y.C.) and the US Department of Energy (A.Y.C.), corporate gifts from Eastman Kodak Company and Peak Performance Technologies (S.L.S.), and research leave from the College of William and Mary (S.L.S.).

Correspondence and requests for materials should be addressed to S.L.S. (e-mail: slsand@wm.edu).

Natural conjugative plasmids induce bacterial biofilm development

Jean-Marc Ghigo

Unité des Membranes Bactériennes Institut Pasteur (CNRS URA 2172),
25 rue du Dr Roux, 75724 Paris Cedex 15, France

Horizontal gene transfer is a principal source of evolution leading to change in the ecological character of bacterial species¹. Bacterial conjugation², which promotes the horizontal transfer of genetic material between donor and recipient cells by physical contact, is a phenomenon of fundamental evolutionary consequence³. Although conjugation has been studied primarily in liquid, most natural bacterial populations are found associated with environmental surfaces in complex multispecies communities called biofilms⁴. Biofilms are ideally suited to the exchange of genetic material of various origins, and it has been shown that bacterial conjugation occurs within biofilms^{5,6}. Here I investigate the direct contribution of conjugative plasmids themselves to the capacity of the bacterial host to form a biofilm. Natural conjugative plasmids expressed factors that induced planktonic bacteria to form or enter biofilm communities, which favour the infectious transfer of the plasmid. This general connection between conjugation and biofilms suggests that medically relevant plasmid-bearing strains are more likely to form a biofilm. This may influence both the chances of biofilm-related infection risks and of conjugational spread of virulence factors.

I used high-debit, continuous flow cultures to monitor biofilm formation of various *Escherichia coli* K12 laboratory strains on submerged, removable Pyrex slides. Strains carrying the F episome,

such as *E. coli* TG1, formed a thick biofilm in 24 h (up to 500 μm with 2×10^{10} colony-forming units (CFU) per cm^2) (Fig. 1a; see also Supplementary Information Fig. 4). In contrast, after 24 h all the F– strains tested only formed microcolonies whose development was not sufficient to be observed macroscopically (8×10^5 CFU cm^{-2} ; Fig. 1a, b). To demonstrate that this phenotype was due to the F factor, I removed the plasmid from TG1 and consequently TG1 F– lost its ability to form a biofilm (Fig. 1a). Furthermore, when conjugated into MG1655, the F plasmid promoted biofilm formation in this and all F– strains tested (Fig. 1c).

The physical nature of conjugation (cell–cell contact through specialized conjugative pili), along with the observation that various pili-like cell appendages are involved in biofilm formation^{7,8}, led me to test the role of conjugative pili produced in *E. coli* carrying the F factor in this biofilm process. I generated a non-polar mutation of the gene that encodes the pilin subunit of the conjugative pilus, *traA*, located within the large *tra* operon⁹ (Fig. 2a). Strain TG1 *traA::aphA* did not form a biofilm and this capacity is restored by complementation with a plasmid bearing the wild-type *traA* allele (Fig. 2b). The same observations were made using transfer-deficient F plasmids indicating that the phenomenon does not require conjugative DNA transfer itself. This result demonstrated the role of the conjugative F pilus in biofilm formation.

The kinetics of surface adhesion in the F– and F+ *E. coli* strains MG1655 and MG1655 F, respectively, showed that expression of the F pilus accelerated both the initial adhesion events and biofilm development (Fig. 1b; see also Supplementary Information Fig. 5). In the culture conditions, biofilm development does not result from recruitment of planktonic cells but from internal growth on the available surfaces. To determine whether pilus-specific donor–recipient interactions could mediate cell contacts, I tested the effect of a mutation in the gene coding for the outer membrane protein OmpA on biofilm growth. OmpA reduces F pilus-specific

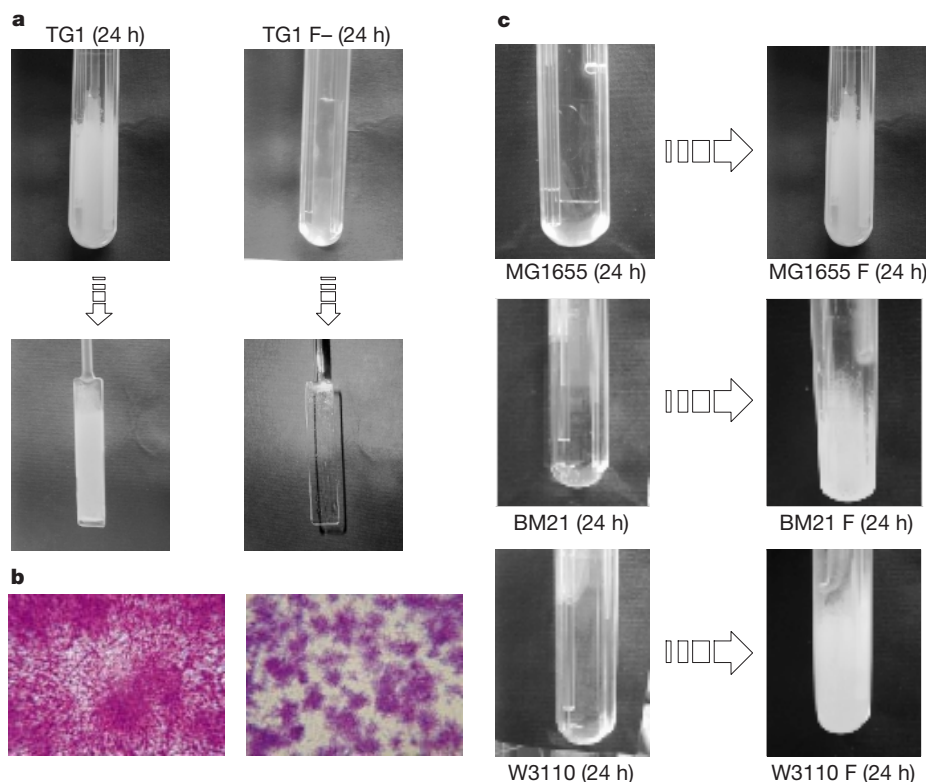


Figure 1 The F episome induces the formation of a thick biofilm in *E. coli*. **a**, Biofilm formation in TG1 and TG1 F– strain. Top, biofilm phenotype in microfermenters. Bottom, biofilm phenotype on the corresponding Pyrex slides removed from the microfermenters. **b**, Comparison of cell density of strain MG1655 F and MG1655 after 24 h of culture in

microfermenters. Transmitted light microscopy of the Pyrex slide stained with crystal violet. Left, MG1655 F biofilm in the area that was least dense. Right, MG1655 biofilm in the area that was most dense. Scale bar, 10 μm . **c**, Biofilm formation in three different *E. coli* K12 strains. Left, F– strains; right, F+ strains.

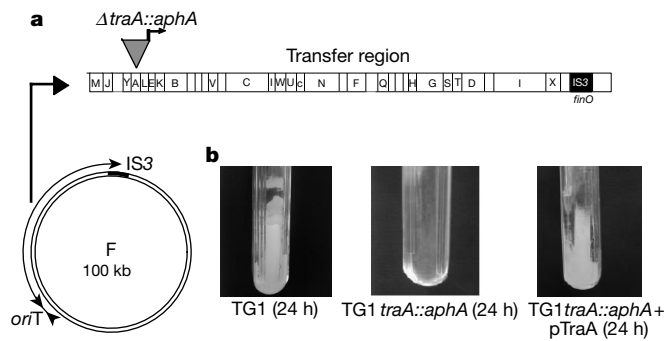


Figure 2 F pilus synthesis is required for biofilm formation in TG1. **a**, Schematic genetic map of the F plasmid's *tra* operon (33.3 kilobases). **b**, Biofilm formation by TG1 after 24 h (left). Biofilm formation by TG1 *traA::aphA* after 24 h (middle). Biofilm formation by TG1 *traA::aphA* complemented with plasmid pTraA after 24 h (right).

contacts with the recipient cell during conjugation¹⁰. No reduction in the extent of biofilm growth could be observed in an F+ *ompA* mutant compared to its wild-type parental isogenic strain. These results suggest that the F conjugative pilus acts as an adhesion factor that allows non-specific cell–surface or cell–cell contacts leading to the observed three-dimensional growth of the biofilm.

I proposed that plasmids isolated from natural populations of Gram-negative bacteria could also promote biofilm formation. Twenty-nine natural conjugative plasmids were tested for their capacity to promote biofilm formation. Out of these 29 plasmids nine from six different incompatibility groups (Table 1; see also

Supplementary Information Fig. 6) led to a biofilm. This demonstrated that induction of biofilm formation is not restricted to a specific incompatibility group nor to a specific type of pili.

These results also showed that plasmids known to be constitutive for pilus synthesis allowed the formation of a biofilm whereas most of the plasmids described as repressed did not (Table 1). Repression results in very few bacteria (0.1%) with conjugative ability, which probably minimizes the burden on the host arising either from the energy expended to maintain the conjugative apparatus or from associated properties, such as pili-specific bacteriophage sensitivity. I investigated the relationship between the regulation of pilus synthesis and the ability to form a biofilm, and compared the repressed F-like natural plasmid, R1, with its permanently derepressed mutant R1-*drd19*. Although MG1655 carrying R1 did not form a biofilm, its mutant R1-*drd19* promoted the formation of a thick biofilm (Fig. 3a). During the conjugative process, chance encounters between rare transfer-proficient bacteria that carry a repressed plasmid and potential recipients initiate a cascade of conjugative transfer because newly transferred episomes are transiently derepressed. This phenomenon, called ‘epidemic spread’, proceeds until all potential recipients have acquired the plasmid¹¹. To further investigate how transitory derepression of the conjugative apparatus could lead to the development of a biofilm, I co-inoculated the donor *E. coli* strain MG1655 R1 with the recipient BM21. This led to an intermediate phenotype between the biofilm capacity of MG1655 R1 alone and MG1655 R1-*drd19* (Fig. 3b). This suggested that the transient derepression of the R1 conjugative functions due to contact with the recipient cells can promote biofilm formation. Such a situation is likely to occur when planktonic bacteria carrying plasmids come into proximity with a biofilm

Table 1 Biofilm formation promoted by natural conjugative plasmids

	Strain	Original host†	Type of pili‡	Pilus synthesis‡	Biofilm <i>per se</i>	Biofilm after co-inoculation
	TG1	NA	Thick flexible	Constitutive	+	NA
	MG1655	NA	none	NA	–	NA
	BM21	NA	none	NA	–	NA
	W3110	NA	none	NA	–	NA
Incompatibility group*	Plasmids†	Original host‡	Type of pili‡	Pilus synthesis	Biofilm <i>per se</i>	Biofilm after co-inoculation
IncFI	F'lac	<i>Escherichia coli</i> K12	Thick flexible	Constitutive	+	ND
	PIP162	<i>Salmonella enterica</i> ser. typhi	Thick flexible	Repressed	–	+
	RIP162-2	<i>Salmonella enterica</i> ser. typhi	Thick flexible	Repressed	–	+
IncFII	R1	<i>Salmonella enterica</i> ser. paratyphi	Thick flexible	Repressed	–	+
	R1drd19	–	–	Constitutive	+	ND
	R1-16	–	–	Repressed	–	+
IncI1	PIP24	<i>Shigella sonnei</i>	–	Repressed	–	+
	pIP112	<i>Salmonella enterica</i> ser. panama	Thin flexible	Repressed	–	+
	R64	<i>Salmonella typhimurium</i>	–	Repressed	–	+
IncJ	R391	<i>Providencia rettgeri</i>	Thick flexible	Repressed	–	+
IncI2	TP114	<i>Escherichia coli</i>	Thin flexible	Repressed	+	ND
	PIP175	<i>Salmonella enterica</i> ser. Wien	Thin flexible	Repressed	–	+
Inc9	RIP71a	<i>Escherichia coli</i>	Thick flexible	Repressed	–	+/-
IncH2	TP116	<i>Salmonella enterica</i> ser. typhi	Thick flexible	Repressed	+	ND
IncX	PIP1100	<i>Escherichia coli</i>	Thick flexible	Repressed	+/-	+
	R6K	<i>Providencia rettgeri</i>	Thick flexible	Repressed	–	+
Inc10	PIP72	<i>Escherichia coli</i>	Thick flexible	Repressed	–	+
IncA/C	R16a	<i>Providencia stuartii</i>	Thick flexible	Repressed	–	+
	pIP55	<i>Klebsiella pneumoniae</i>	Thick flexible	Repressed	+/-	+/-
IncT	pRtsI	<i>Proteus vulgaris</i>	Thick flexible	Repressed	+/-	+
Inc7/L/M	pIP135	<i>Enterobacter cloacae</i>	Rigid	Repressed	+	ND
	pIP69	<i>Salmonella enterica</i> ser. paratyphi	Rigid	Repressed	–	+/-
	R69-2	–	Rigid	Repressed	+	ND
	pIP113	<i>Salmonella enterica</i> ser. panama	Rigid	Repressed	+	ND
IncP	RN3	<i>Shigella</i> sp.	Rigid	Constitutive	+	ND
	RP4	<i>Pseudomonas aeruginosa</i>	Rigid	Repressed	–	+
	RP1	<i>Pseudomonas aeruginosa</i>	Rigid	Repressed	–	+
IncW	R702	<i>Proteus mirabilis</i>	Rigid	Repressed	–	+
	RSa	<i>Shigella</i> sp.	Rigid	Constitutive	+	ND
	R388	<i>Escherichia coli</i>	Riaid	Constitutive	+	ND

NA, not applicable; ND, not determined; +, formation of a thick biofilm (2.1×10^8 CFU cm⁻²); +/-, formation of a weak but visible biofilm (5.1×10^7 CFU cm⁻²); –, no biofilm macroscopically observed ($<10^5$ to 10^7 CFU cm⁻²).

* Data from ref. 16.

† Data from ref. 17.

‡ Data from refs. 18 and 19.

population composed of potential recipient cells, that is, bacteria that do not carry conjugative plasmids of the same incompatibility group.

To mimic a situation where donor bacteria carrying repressed conjugative plasmid(s) would encounter a surface-attached recipient population, I first inoculated BM21 for 24 h. These bacteria reached the surface and formed microcolonies without much further development. After 24 h, the co-inoculation with MG1655 R1 led to the formation of a biofilm similar in extent to the one formed by MG1655 R1-*drd19*, where up to 90% of the cells corresponded to recipient bacteria that received R1, less than 2% to donor bacteria (MG1655 R1) and up to 10% of plasmid-free, initial recipient cells (BM21). This suggests that derepression leads to both biofilm formation and horizontal transfer of the plasmid to the recipient cells. In a reverse experiment where the donor MG1655 R1 was inoculated 24 h before the recipient BM21, both the extent of transfer and biofilm development were reduced (Fig. 3b). This may be a consequence of issues concerning accessibility, where surface-attached MG1655 R1 are not in a position to interact efficiently with a high enough number of planktonic recipient cells.

I performed co-inoculation experiments with the 20 strains carrying plasmids that do not promote a thick biofilm formation on their own (Table 1). When the plasmid-free MG1655 strain was inoculated 24 h before the plasmid-carrying strains, 17 out of 20 of these strains produced a thick biofilm (Table 1; see also Supplementary Information Fig. 7). Thus, 90% (26) of the 29 natural conjugative plasmids that were tested induced the development of a biofilm either alone or in the presence of recipient cells.

I propose that biofilm formation proceeds through the introduction of a small sub-population of donor bacteria expressing the conjugative pili into the surface-attached recipient population. After this initial event, the conjugative transfer of the plasmid occurs and the conjugative pili synthesis is derepressed in the transconjugants and their immediate descendants. The recipient population transiently expressing conjugative pili would then form a biofilm through pilus-mediated networking within the fast-growing upper layer bacteria of the biofilm (see Fig. 3c). On completion of the plasmid transfer to the surface-attached bacteria, the conjugative pili would recede and be replaced by adhesion

factors (such as other types of pili or exopolysaccharides), maintaining the coherence of the biofilm.

I have shown here that conjugative plasmids express factors that favour the access of planktonic bacteria to biofilm communities. By doing so, they increase their chances of being transferred to a large number of recipient cells, and support the emerging hypothesis that biofilms may constitute ecological niches that are highly favourable to conjugation¹².

The persistence of nearly ubiquitous plasmids in natural bacterial populations requires one or both of two basic mechanisms: parasitic transmission through horizontal transfer or vertical transmission through maintenance in the host progeny. Empirical studies and mathematical models suggest that plasmids cannot persist solely by carrying genes that are beneficial to their bacterial host¹³. On the other hand, it has been proposed that plasmids cannot be maintained as pure genetic parasites owing to their low rate of infectious transmission, which cannot overcome the counter selection due to the burden that their replication imposes on their host. The transmission of bacterial plasmids has therefore often been considered to be a trade off between horizontal and vertical modes of transmission. However, these assumptions were made on the basis of a limited number of laboratory studies essentially performed in liquid culture, where transfer dynamics may not reflect those of natural bacterial populations such as biofilms¹³.

I propose that conjugative plasmids favour the incorporation of their host into biofilm environments where the density of potential bacterial recipients is sufficient for plasmids to transfer at high rates and to be maintained by infectious transmission alone within some natural habitats. Hence, the expression of biofilm-promoting factors, without having an *a priori* benefit for their host, could sufficiently enhance the probability of the infectious transfer rate so as to offset the segregation loss.

Joining a biofilm community may also have numerous beneficial aspects for the bacterial host itself, such as collective defence against physical and chemical environmental stresses¹⁴. Such ecologically beneficial factors expressed by plasmids may provide a rationale for the unexplained vertical maintenance of the numerous uninfected cryptic plasmids found in natural bacterial populations.

Conjugative plasmids mediate the horizontal transfer of anti-

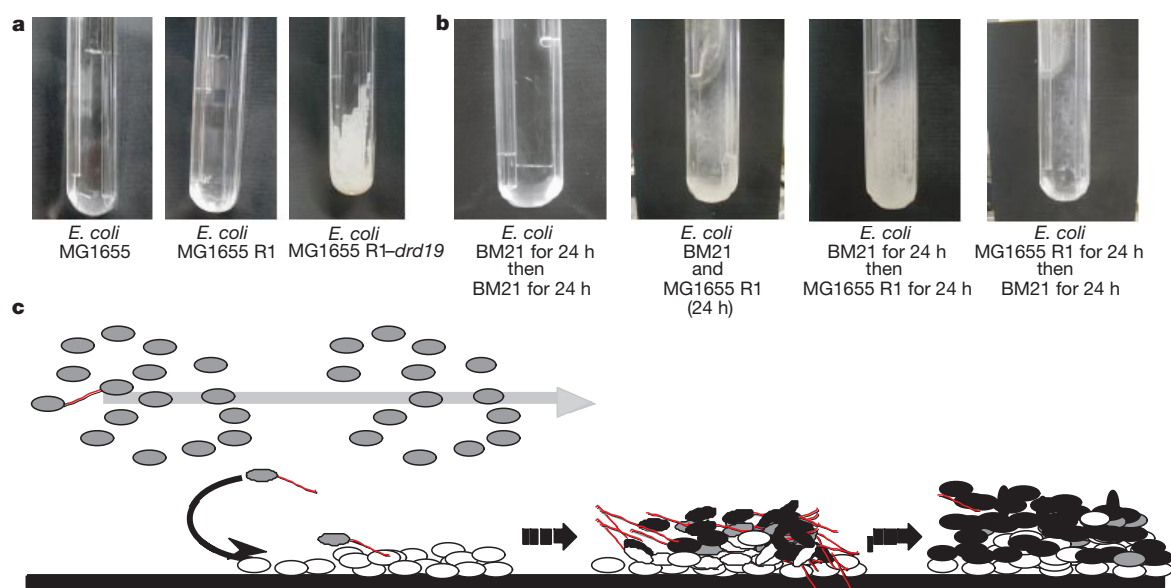


Figure 3 Derepression of conjugation ability induces biofilm development. **a**, Biofilm formation comparison between plasmid-free MG1655 strain (left), and MG1655 carrying the repressed conjugative plasmid R1 (middle) or its derepressed mutant R1-*drd19* (right). **b**, Co-inoculation experiments. **c**, Schematic model. Grey planktonic cells indicate

donor cells carrying the conjugative plasmids, which express a biofilm-promoting factor (for example, conjugative pilus in red). White cells indicate recipient cells and black cells indicate transconjugants (see text).

biotic resistance and virulence factors. The finding of a connection between bacterial conjugation and biofilm formation suggests that an important ecological consequence of the use of antibiotics and biocides in clinical medicine and agriculture may have been the selection of plasmid-bearing strains that are more likely to form a biofilm. Because biofilms are a common cause of persistent nosocomial infections that are difficult to eradicate owing to innate physiological properties¹⁴, this aspect may prove to be of relevant medical significance in addition to the conjugational spread of virulence factors themselves. □

Methods

Bacterial strains and plasmids

Bacterial strains are listed in Table 1 and were provided by the Collection of the Institut Pasteur (<http://www.pasteur.fr/applications/CIP/>) and the Unité des Agents Antibactériens (P. Courvalin and G. Gerbaud).

Biofilm

All experiments were performed in triplicate in 0.4% glucose M63B1 minimal medium at 37 °C. Continuous 60-ml microfermenters with four liquid and gas sampling ports (Pasteur Institute's Laboratory of Fermentation) were configured as continuous-flow culture bioreactors with a 40 ml h⁻¹ flow rate (*F*). 10⁸ bacterial inocula from overnight precultures grown in glucose minimal medium with required antibiotics were used to inoculate microfermenters, which were then cultivated for 3–48 h. The culture volume (*V*) was constant and the imposed dilution rate (*D*) was $D = F/V = 0.66 \text{ h}^{-1}$. Hence, the theoretical generation time (*T*) required for constant density culture in the micro-fermenter was $T = \ln 2/D = 1.05 \text{ h}$. The average generation time calculated in exponential batch culture for *E. coli* strains MG1655 and BM21 was 1.3 h. Therefore, the high input rate of fresh, diluting medium used in our experimental model was imposed to avoid any significant planktonic growth. Stirring was assured by aeration with sterile pressed air (0.3 bar). Submerged, removable Pyrex slides (total area of 22.4 cm²) served as growth substratum.

Microscopy and image analysis

Biofilm development was recorded with a Nikon Coolpix 950 digital camera. Epifluorescence, phase contrast and transmitted light microscopy were acquired with a Leitz Dialux 20EB microscope equipped with $\times 25$ to $\times 100$ objectives. Scanning confocal microscopy was performed at the confocal microscopy station of the Pasteur Institute.

Non-polar deletion of the *traA* gene

A non-polar mutation deleting the entire *traA* gene was created by allelic exchange¹⁵ using the primers TraAGB-5': 5'-AGGGAGGCAGATAAAGAGGAAGATATAACATTTAATACA CTCTAGTTTATTCATTTATCCGAAATTGAGGTAACCTATGAAAGCCACGTTGTG TCTCAA-3' and TraAGBnp-3': 5'-GCGGCTCTGGTTGGTCAGTGTTCGCGGAAACG ATATTCTTAAGTTTATTCCTGCTCTCCGACATCGTTTATTTCTCTGTAGAAAAA CTCATCGAGCA-3', and *aphA* gene (kanamycin resistance) from Tn903 as template. The mutation was verified by PCR analysis.

Cloning of the *traA* gene

pTraA was constructed by PCR amplification of *traA* from strain TG1 using the primers TraAecorbs-5': 5'-AAAGAATTCGAAATTGAGGTAACCTATGAATGC-3' and TraAH3-3': 5'-CCCAAGCTTCGTTTATTTCTCTGTACAG-3'. I verified the nucleotide sequence of the construction.

Biofilm co-inoculation procedures

A preculture of a recipient strain BM21 (nal^r (nalidixic acid resistant)) was inoculated for 24 h and then co-inoculated with MG1655-S R1 (St^r (streptomycin resistant), Ap^r (ampicillin resistant) and Km^r (kanamycin resistant)) for another 24 h. Pyrex slides were removed and centrifuged in 15 ml of fresh M63B1 medium for 1 min. The number of colony-forming units was determined by plating serial dilutions of the resuspensions on medium supplemented with nalidixic acid (for recipient BM21 scoring), streptomycin, ampicillin, kanamycin (for donor MG1655-S R1 scoring), nalidixic acid, ampicillin and kanamycin (for BM21 R1 transconjugant scoring), and without antibiotic (for total cell scoring). Co-inoculation with BM21 strains carrying the plasmids described in Table 1 were generated using MG1655-S as recipient bacteria.

Received 8 March; accepted 6 June 2001.

- de la Cruz, I. & Davies, I. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8**, 128–133 (2000).
- Lederberg, J. & Tatum, E. L. Gene recombination in *Escherichia coli*. *Nature* **158**, 558 (1946).
- Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
- Costerton, J. W., Lewandowski, Z., Caldwell, D. E., Korber, D. R. & Lappin-Scott, H. M. Microbial biofilms. *Annu. Rev. Microbiol.* **49**, 711–745 (1995).
- Hausner, M. & Wuertz, S. High rates of conjugation in bacterial biofilms as determined by quantitative in situ analysis. *Appl. Environ. Microbiol.* **65**, 3710–3713 (1999).

- Christensen, B. B. *et al.* Establishment of new genetic traits in a microbial biofilm community. *Appl. Environ. Microbiol.* **64**, 2247–2255 (1998).
- O'Toole, G. A. & Kolter, R. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol.* **30**, 295–304 (1998).
- Pratt, L. A. & Kolter, R. Genetic analyses of bacterial biofilm formation. *Curr. Opin. Microbiol.* **2**, 598–603 (1999).
- Frost, L. S., Ippen-Ihler, K. & Skurray, R. A. Analysis of the sequence and gene products of the transfer region of the F sex factor. *Microbiol. Rev.* **58**, 162–210 (1994).
- Ried, G. & Henning, U. A unique amino acid substitution in the outer membrane protein OmpA causes conjugation deficiency in *Escherichia coli* K-12. *FEBS Lett.* **223**, 387–390 (1987).
- Lundquist, P. D. & Levin, B. R. Transitory derepression and the maintenance of conjugative plasmids. *Genetics* **113**, 483–497 (1986).
- Watnick, P. & Kolter, R. Biofilm, city of microbes. *J. Bacteriol.* **182**, 2675–2679 (2000).
- Bergstrom, C. T., Lipsitch, M. & Levin, B. R. Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics* **155**, 1505–1519 (2000).
- Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284**, 1318–1322 (1999).
- Chaveroche, M. K., Ghigo, J. M. & d'Enfert, C. A rapid method for efficient gene replacement in the filamentous fungus *aspergillus nidulans*. *Nucleic Acids Res.* **28**, E97 (2000).
- Couturier, M., Bex, F., Bergquist, P. L. & Maas, W. K. Identification and classification of bacterial plasmids. *Microbiol. Rev.* **52**, 375–395 (1988).
- Bukhari, A. L., Shapiro, J. A. & Adhya, S. L. (eds) *DNA Insertion Elements, Plasmids and Episomes* 601–656 (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1977).
- Bradley, D. E., Taylor, D. E. & Cohen, D. R. Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12. *J. Bacteriol.* **143**, 1466–1460 (1980).
- Frost, L. S. in *Bacterial conjugation* (ed. Clewley, D. B.) 189–221 (Plenum, New York, 1993).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

I am grateful to C. Wandersman for support and constant interest during the course of this work. I also thank D. Mazel, G. Gerbaud and P. Courvalin for providing the plasmids and strains used in this study; A. Idja and P. Roux for technical assistance; R. Longin for providing facilities of the Pasteur Institute Laboratory of Fermentations; and D. Mazel, E. Stewart, D. A. Rowe-Magnus and P. Delepleaire for helpful discussions and critical reading of the manuscript. This work was supported by grants from the Programme de Recherche Fondamentale en Microbiologie et Maladie Infectieuses, réseau Infections Nosocomiales (MENRT) and the Pasteur Institute.

Correspondence and requests for material should be addressed to the author (e-mail: jmghigo@pasteur.fr).

Force can overcome object geometry in the perception of shape through active touch

Gabriel Robles-De-La-Torre & Vincent Hayward

McGill University, Center for Intelligent Machines, Montréal, Canada H3A 2A7

Haptic (touch) perception normally entails an active exploration of object surfaces over time. This is called active touch^{1–3}. When exploring the shape of an object, we experience both geometrical⁴ and force cues. For example, when sliding a finger across a surface with a rigid bump on it, the finger moves over the bump while being opposed by a force whose direction and magnitude are related to the slope of the bump⁵. The steeper the bump, the stronger the resistance. Geometrical and force cues are correlated, but it has been commonly assumed that shape perception relies on object geometry alone. Here we show that regardless of surface geometry, subjects identified and located shape features on the basis of force cues or their correlates. Using paradoxical stimuli, for example combining the force cues of a bump with the geometry of a hole, we found that subjects perceived a bump. Conversely, when combining the force cues of a hole with the geometry of a bump, subjects typically perceived a hole. In two experiments, human subjects explored surfaces by touch

using an apparatus that allowed us to separate force cues from surface geometry. Subjects explored the shape of surfaces through a manipulandum placed behind a curtain (Fig. 1). Subjects pressed down on the plate of the manipulandum with their index fingers while smoothly rolling it on the surfaces. There were three interchangeable physical surfaces. One surface was flat, one had a bump, and another had a hole; the latter two had a gaussian profile that was 0.3 cm high/deep and 4 cm wide. The plate was constrained to remain horizontal, and its vertical position was entirely determined by the geometry of the physical surfaces. A force-feedback haptic interface could produce a horizontal force as a function of the measured vertical component F_{sy} (the force applied vertically by the subjects, Fig. 2a) and of the plate's horizontal position (Fig. 1, Methods). The force of the interface interacted with the force returned by the physical surface. The interface's force was called a 'virtual surface' because it provided the same horizontal force component that an equivalent physical surface would return, regardless of the manipulandum's vertical position. For example, when subjects explored a physical bump or hole (Fig. 2a and b), they experienced a horizontal force F_{px} . When subjects explored a flat surface combined with a virtual bump or hole, they experienced the same horizontal force F_{px} (Fig. 2c, grey, dashed curve), but the manipulandum moved in a straight line (Fig. 2c, black line). In another example, the horizontal force component from a physical bump was cancelled out ('force-masked', Methods) by a spatially aligned virtual hole (Fig. 2d). Here, the vertical position of the manipulandum followed the geometry of the bump (Fig. 2d, black curve), but subjects did not experience horizontal force components (Fig. 2d, grey, dashed line). The manipulandum would stay in equilibrium on a slope regardless of how much the subjects pushed on it, just as if they were exploring a flat surface.

In experiment 1, we used the paradoxical stimuli just described to investigate whether subjects classified and located shape features by following geometrical or horizontal force cues. Subjects were tested using seven different conditions. In condition 1, only flat surfaces were presented. In condition 2, the flat surfaces were combined with virtual bumps, and in condition 3 with virtual holes (Fig. 2c). In condition 4 there were only ordinary physical bumps, and in condition 5 ordinary physical holes (Fig. 2b). In condition 6, the physical bumps were force-masked by virtual holes, and control virtual bumps were randomly placed elsewhere (Fig. 2e). Condition 7 mirrored condition 6 (Fig. 2e). The positions of physical and virtual surfaces were uncorrelated in conditions 6 and 7. The probability of classifying a stimulus as a hole or a bump was calculated for each subject under all conditions. The stimulus localization performance of the subjects was measured by the correlation between stimulus location and subjects' positioning of the manipulandum (Methods).

The flat surfaces were classified as flat (Fig. 3a, b, condition 1).

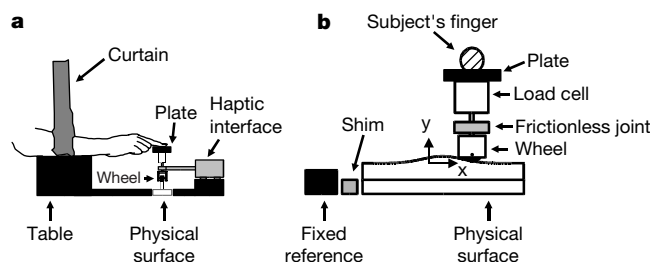


Figure 1 Side (a) and front (b) views of the apparatus. Subjects pressed down on the manipulandum's plate and rolled it sideways (x -axis in b) to explore the shape of an interchangeable physical surface. The entire workspace was hidden from view behind a black curtain. The manipulandum was connected to a haptic interface (Methods) that operated silently.

Virtual bumps, combined with the flat surface, were classified as bumps ($P < 0.001$; Fig. 3a, condition 2), and were accurately located (Fig. 3c). Virtual holes, combined with the flat surface, were also identified as holes ($P < 0.001$; Fig. 3b, condition 3) and were precisely located (Fig. 3d). When physical holes or bumps were presented alone, subjects also easily identified and located them (Fig. 3a–d, conditions 4 and 5). There was no significant difference in subjects' identification and localization performance when exploring physical or virtual surfaces. However, when force-masked physical holes or bumps were combined with virtual surfaces, subjects' features localization was affected drastically. Subjects' identification performance remained the same (Fig. 3a, b, conditions 6 and 7), but most subjects tracked the control virtual

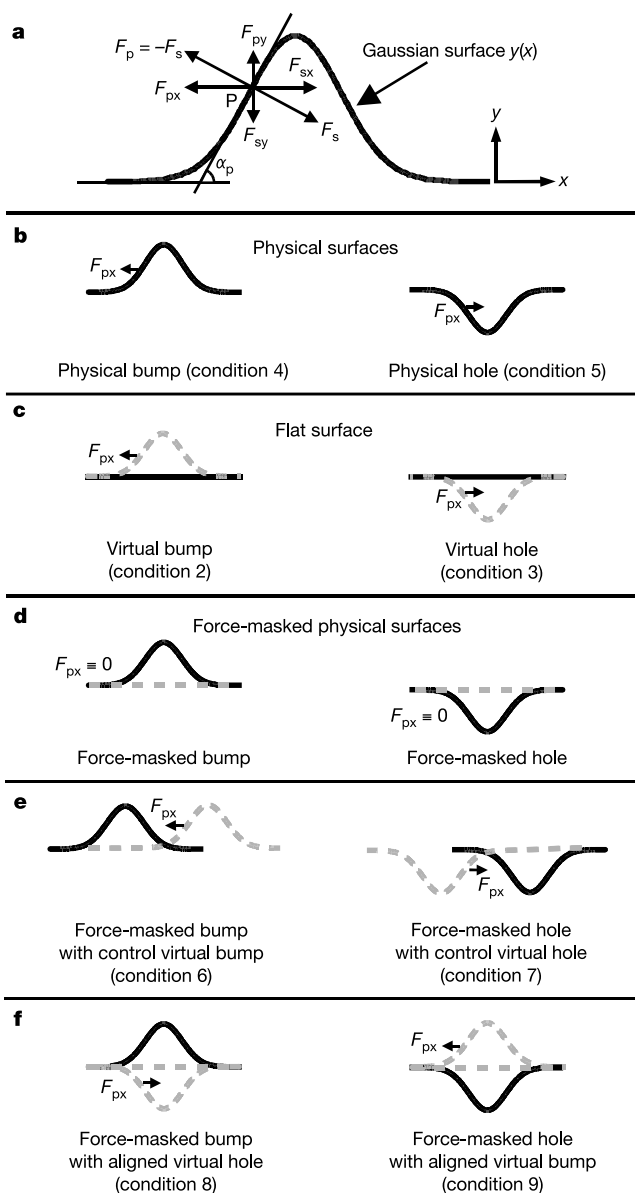


Figure 2 Experimental conditions 2–9. When the haptic interface was turned off, subjects experienced F_{px} , the horizontal force component that depends on surface geometry (a, b). Equivalent forces (c–f) were produced by the interface to generate stimuli with uncorrelated force (grey, dashed curves) and geometric cues (black curves). The force cues from virtual bump/hole stimuli (c) were equivalent to those of physical bumps/holes, but provided the geometrical cues of a flat surface. Force-masked stimuli (d–f) were used to create physical bumps/holes with spatially uncorrelated cues (e), physical holes with a bump's force cues, and physical bumps with a hole's force cues (f).

surfaces instead of the force-masked physical holes or bumps (Fig. 3c, d, conditions 6 and 7, thick circles, $P < 0.009$ for virtual bumps and $P < 0.001$ for virtual holes; see also Supplementary Information). The physical bumps or holes provided geometric but not horizontal force information to subjects, whereas the virtual surfaces provided force but not geometrical information. Given the subjects' instructions (see Methods), this suggests that the control virtual holes or bumps seemed deeper or bigger than the force-masked physical holes or bumps. Horizontal force was used instead of geometry to identify and locate shape features. This happened when geometrical information was absent (conditions 2 and 3) and when object geometry did not correlate with force (conditions 6 and 7). One subject displayed a significant localization performance when tracking force-masked bumps (Fig. 3c, condition 6). The position of the force-masked and virtual bumps presented to this subject had a spurious, non-significant correlation of 0.32 that explained the subject's significant tracking. Only this subject was exposed to spurious correlations. However, the subject clearly followed the virtual bumps and not the force-masked physical bumps (Supplementary Information). Some subjects located the force-masked shapes in some trials, and the control virtual shapes in other trials. These were subjects with significant localization correlations equal to or below 0.65 when tracking virtual bumps (Fig. 3c, condition 6; Supplementary Information), or virtual and force-masked holes (Fig. 3d, condition 7; Supplementary Information).

If subjects relied on force information to identify and locate shape features, then an easily identified and located physical bump should be perceptually transformed into a hole by combining it with the force cues of a hole. An equivalent perceptual reversal should happen for a physical hole. Experiment 2 explored this idea with a new group of subjects. Again, they were tested using condition 1 (flat surfaces, no virtual surface), conditions 2 and 3 (virtual surfaces, Fig. 2c), and conditions 4 and 5 (physical surface alone,

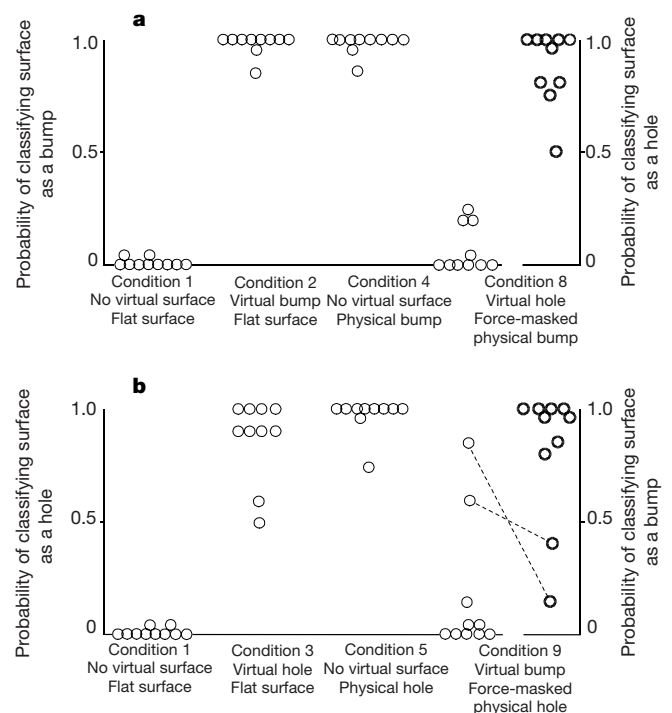


Figure 4 Force perceptually transformed geometrical shape. Each symbol represents data from a single subject. Subjects easily classified physical and virtual surfaces (a, b, conditions 1–5) as bumps or holes. However, when force-masked physical bumps coincided with virtual holes, stimuli were classified as holes (a, condition 8, thick circles, right axis). A mirrored perceptual change occurred when force-masked physical holes coincided with virtual bumps (b, condition 9, thick circles, right axis). Two subjects identified the physical holes but took the longest average times to make a judgement (text). Dotted lines join data points corresponding to the same subjects.

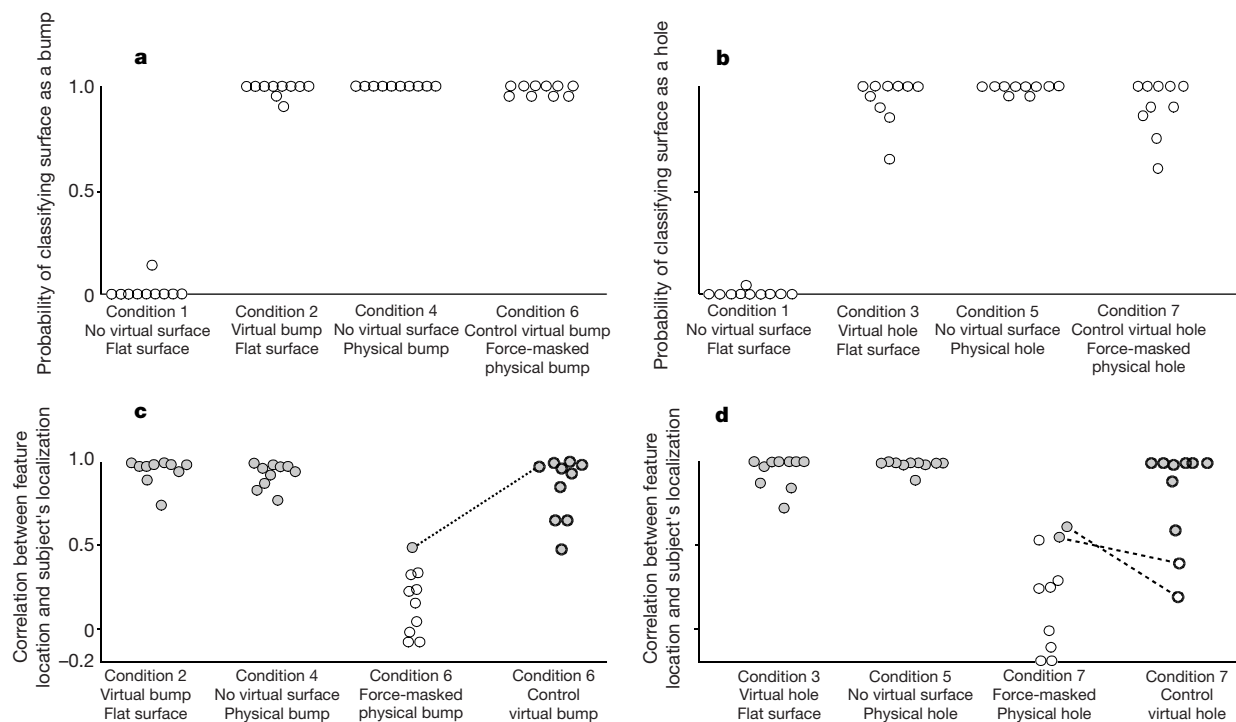


Figure 3 Subjects used force to identify or locate shape features. Each symbol represents data from one subject. Subjects easily classified stimuli as bumps or holes when a virtual surface or flat physical surface combination was presented, or when virtual and physical surfaces coexisted, but not when a flat surface was presented alone (a, b). Subjects tracked the position of physical and virtual surfaces (c, d, conditions 2–5). However,

when virtual surfaces and force-masked physical surfaces were presented side by side, most subjects tracked the control virtual surfaces (c, d, conditions 6 and 7, thick circles). Dotted lines link the data from the same subject. Grey circles indicate significant correlations ($P < 0.05$) between stimulus position and subjects' localization.

Fig. 2b). Conditions 6 and 7 were replaced by conditions 8 and 9. In condition 8, two virtual holes coincided with a physical bump (Fig. 2f). One virtual hole force-masked the physical bump, and the other provided a hole's horizontal force cues. As a result, the net horizontal force helped subjects to ascend slopes. Condition 9 mirrored condition 8 (Fig. 2f).

Subjects easily identified or located physical bumps when there was no virtual surface (Fig. 4a, condition 4). However, when physical bumps were presented together with the force cues of a hole, subjects reversed their identification performance. These stimuli were not classified as bumps (Fig. 4a, condition 8, thin circles, left axis) but as holes ($P < 0.001$, Fig. 4a, condition 8, thick circles, right axis). One subject identified these stimuli as either flat surfaces or holes with a 50% probability. Physical holes were easily identified or located when presented alone (Fig. 4b, condition 5). But combinations of physical holes and virtual bumps were classified as bumps ($P < 0.009$, Fig. 4b, condition 9, thick circles, right axis), not as holes (Fig. 4b, condition 9, thin circles, left axis). Two subjects behaved differently (Fig. 4b, data points joined by dotted lines). These two subjects took the longest average time to explore these stimuli: 20.4 s and 21 s, compared to 6.8–16.7 s (mean 11.4 s) for all other subjects. This suggested that these stimuli were ambiguous to subjects and that, over longer exploration times, geometrical cues may contribute more to subjects' perception. As in experiment 1, there was no significant difference in subjects' ability to locate any of the stimuli categories (data not shown).

Our results indicate that force or force-related information (for example, hand/finger velocity) can overcome geometrical information, such as slope, to determine shape perception through active touch. This happened for maximum slope differences well above reported thresholds: 5.71 degrees compared to 0.57 (ref. 6) and 3–4 degrees (ref. 4). Our subjects used a combination of finger, wrist and forearm movements to explore stimuli. Kinesthesia research^{7–9} suggests that subjects may have easily detected these movements, particularly those of the metacarpophalangeal joint of the index finger⁹. Our results indicate that most subjects did not use this information to classify/locate shape features. This suggests that the nervous system may sense both force and kinesthetic cues for shape perception, but processes these cues separately and/or weighs them differently.

Because classification or localization performance for physical and virtual surfaces was equivalent, virtual surfaces can be considered as illusory haptic shapes. Informal observations suggest that similar phenomena occur when the manipulandum is used with the wrist, the elbow and the big toe. The reader may reproduce the approximate conditions of our experiments with simple materials (see <http://www.cim.mcgill.ca/~roblesg/vsdemo.html>). □

Methods

Physical shapes

When subjects explored a frictionless physical surface $y(x)$, they applied a force F_x (Fig. 2a). The surface returned a normal force $F_y = -F_x$ at P . F_{px} and F_{py} (the horizontal and vertical components of F_p) are related: $F_{px} = F_{py} \tan[\alpha_p(x)]$. For a profile $y(x)$, $\tan[\alpha_p(x)] = dy/dx$. The profile used was gaussian, $y(x) = k \exp[-(x - x_{ploc})^2/w^2]$, where x_{ploc} is the position of the profile in the workspace, $w = 2$ cm, $k = 0.3$ cm for a bump and $k = -0.3$ cm for a hole. Interchangeable shims could precisely locate the profile at $x_{ploc} = 0, 1.56$ or 2.2 cm. The physical surfaces were precision-machined out of hard plastic.

Virtual shapes and force masking

A force-feedback haptic interface produced the computer-controlled horizontal force F_v that interacted with F_x and F_p through a frictionless point (Fig. 1). Under low-velocity conditions, F_v , F_x and F_p balanced: $F_x + F_p + F_v = 0$. Along the x and y directions, $F_{xx} + F_{px} + F_v = 0$ and $F_{yy} + F_{py} = 0$. A load cell (Fig. 1) measured $F_{py} = -F_{px}$ (Fig. 2a) and the haptic interface measured x , the horizontal position of the plate. The device was programmed to produce $F_v = F_{py} \tan[\alpha_p(x - x_{ploc})] + F_{py} \tan[\alpha_v(x - x_{vloc})]$. Because the force experienced by a subject in the horizontal direction was $F_{xx} = -F_{px} - F_v$, the first term of F_v defined a virtual surface that cancelled out ('force-masked') the horizontal component of the force returned by a physical surface located at x_{ploc} . The second term of

F_v defined a virtual surface located at x_{vloc} . The physical/virtual surface shapes were aligned when $x_{ploc} = x_{vloc}$. In conditions 2 and 3 (both experiments), the first term of F_v was not used.

Force-feedback haptic interface

A PenCAT/Pro (Immersion Canada Inc.) interface produced F_v and measured x . The interface had negligible friction and operated silently. Manipulandum's position/forces were sampled at 1 kHz. The force feedback was updated at the same rate. A manipulandum similar to ours was theoretically described by Minsky¹⁰ to illustrate a physical model of simulated textures.

Experimental procedure

Subjects gave informed consent, were paid to participate and naive to the purpose of the experiment, did not have calluses on their right index finger nor report any hand injury/disease. Ten right-handed subjects participated in experiment 1; four males and six females, ages 19–31. Ten new right-handed subjects participated in experiment 2; six males and four females, ages 18–31. Handedness was evaluated by using a standard questionnaire¹¹. In each trial, a stimulus was randomly selected from an experimental condition. The manipulandum's position was randomly set either to the right or left end of the workspace. A subject's index finger was placed on the centre of the manipulandum's plate (Fig. 1). Subjects were instructed to locate the highest/lowest point of the highest/deepest perceived bump/hole. There was no time limit for exploration, but proceeding quickly was encouraged. After locating the feature, subjects did not change the manipulandum's position. Subjects ended the trial by pressing a button on a computer keyboard to identify the located feature. Buttons were labelled 'bump', 'hole' and 'flat'. No feedback was given. Subjects withdrew their fingers from the manipulandum after pressing the button. Only physical surfaces were presented during 25 practice trials. During practice, the shims were not used and the surface's position was randomly varied across the workspace. Each subject proceeded to complete 140 experimental trials. Each experimental condition was tested 20 times. Subjects had periodic breaks after every 50 trials, but could rest at any time. A typical experiment lasted about one hour and fifty minutes.

Data analysis

For each experimental condition, subjects' localization performance was measured by the Pearson correlation coefficient between physical/virtual surface location and subjects' final manipulandum position. If a stimulus provided the force cues of a bump (virtual or physical) but was classified as a hole or as a flat surface, the trial was not used to compute the correlation. Conversely, a trial was not used for correlation calculations if the stimulus provided the force cues of a hole (virtual or physical), but was classified as a bump or as a flat surface. Subjects' classification performance was measured by calculating the probabilities of classifying a stimulus as a bump or as a hole, respectively. Multivariate ANOVA tests and planned comparisons evaluated within-subject effects of each condition.

Received 28 November 2000; accepted 17 May 2001.

- Gibson, J. J. Observations on active touch. *Psychol. Rev.* **69**, 477–491 (1962).
- Loomis, J. M. & Lederman, S. J. in *Handbook of Perception and Human Performance* Vol. 2 *Cognitive Processes and Performance* (ed. Boff, K. R. et al.) Ch. 31, 1–41 (New York, Wiley, 1986).
- Lederman, S. J. & Klatzky, R. L. Hand movements: a window into haptic object recognition. *Cogn. Psychol.* **19**, 342–368 (1987).
- Pont, S. C., Kappers, A. M. L. & Koenderink, J. J. Similar mechanisms underlie curvature comparison by static and dynamic touch. *Percept. Psychophys.* **61**, 874–894 (1999).
- Robles-De-La-Torre, G. & Hayward, V. in *Proc. ASME Dynamics Systems and Control Division* Vol. 2 (ed. Nair, S. S.) 1081–1085 (The American Society of Mechanical Engineers, New York, 2000).
- Gordon, I. E. & Morison, V. The haptic perception of curvature. *Percept. Psychophys.* **31**, 446–450 (1982).
- Taylor, J. L. & McCloskey, D. I. Detection of slow movements imposed at the elbow during active flexion in man. *J. Physiol.* **457**, 503–513 (1992).
- Hall, L. A. & McCloskey, D. I. Detections of movements imposed on finger, elbow and shoulder joints. *J. Physiol.* **335**, 519–533 (1983).
- Clark, F. J., Burgess, R. C. & Chapin, J. W. Proprioception with the proximal interphalangeal joint of the index finger. Evidence for a movement sense without a static-position sense. *Brain* **109**, 1195–1208 (1986).
- Minsky, M. *Computational Haptics: The Sandpaper System for Synthesizing Texture for a Force-Feedback Display*. PhD dissertation. Massachusetts Institute of Technology (1995).
- Coren, S. *The Left-hander Syndrome* (Maxwell Macmillan International, New York, 1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Cynader, B. Frost and L. Requa for helpful comments, and D. Pavlasek and J. Boka for help in designing/building the manipulandum and surfaces. The research was funded by Canada's Network of Centers of Excellence programme, Institute for Robotics and Intelligent Systems, and the Natural Sciences and Engineering Research Council of Canada.

Correspondence and requests for materials should be addressed to G.R. (e-mail: roblesg@cim.mcgill.ca).

Ubiquitination-dependent mechanisms regulate synaptic growth and function

Aaron DiAntonio*, Ali P. Haghighi†, Scott L. Portman*, Jason D. Lee†, Andrew M. Amaranto* & Corey S. Goodman†

* Department of Molecular Biology and Pharmacology, Washington University School of Medicine, 660 S. Euclid, Campus Box 8103, St Louis, Missouri 63110, USA

† Department of Molecular and Cell Biology, University of California, Berkeley, Room 509 Life Sciences Addition, Berkeley, California 94720, USA

The covalent attachment of ubiquitin to cellular proteins is a powerful mechanism for controlling protein activity and localization¹. Ubiquitination is a reversible modification promoted by ubiquitin ligases and antagonized by deubiquitinating proteases². Ubiquitin-dependent mechanisms regulate many important processes including cell-cycle progression, apoptosis and transcriptional regulation³. Here we show that ubiquitin-dependent mechanisms regulate synaptic development at the *Drosophila* neuromuscular junction (NMJ). Neuronal overexpression of the deubiquitinating protease *fat facets*⁴ leads to a profound disruption of synaptic growth control; there is a large increase in the number of synaptic boutons, an elaboration of the synaptic branching pattern, and a disruption of synaptic function. Antagonizing the ubiquitination pathway in neurons by expression of the yeast deubiquitinating protease UBP2 (ref. 5) also produces synaptic overgrowth and dysfunction. Genetic interactions between *fat facets* and *highwire*⁶, a negative regulator of synaptic growth that has structural homology to a family of ubiquitin ligases, suggest that synaptic development may be controlled by the balance between positive and negative regulators of ubiquitination.

Synaptic morphology is dynamic; once formed, synapses expand, retract, and remodel throughout life. This plasticity underlies the refinement of neuronal circuits during development and may be critical for plasticity in the adult brain. To identify molecular mechanisms regulating the morphological growth of synapses, we performed a genetic screen (see Methods) for molecules whose neuronal overexpression disrupts synaptic growth control at the *Drosophila* NMJ. We screened a collection of flies capable of the targeted overexpression of endogenous *Drosophila* genes⁷ and identified two lines, EP(3)381 and EP(3)3520, whose overexpression in the nervous system leads to synaptic overgrowth (Fig. 1). We found that both EP(3)381 and EP(3)3520 overexpress *fat facets* (*faf*), a deubiquitinating protease⁴ (Fig. 1a). Endogenous *faf* transcript is strongly and widely expressed in the developing central nervous system (CNS) (Fig. 1a, b), demonstrating that neuronal expression of *faf* from the EP lines produces overexpression, not misexpression, of the transcript.

Anatomical analysis at the NMJ reveals that neuronal expression from both EP(3)381 and EP(3)3520 leads to an increase both in the number of synaptic boutons and in the synaptic span (the extent of the muscle covered by the synapse) (Fig. 1c–f). This increase is not seen in flies that do not overexpress *faf* (*elav-Gal4* crossed to +) or that overexpress a non-functional *faf* gene (*elav-Gal4* crossed to EP(3)381^{faf}) (Fig. 1g, h). Neuronal overexpression of *faf* also causes an increase in the number of synaptic branches as quantified by the number of branch points (muscle 12: *elav-Gal4/+*, 6.4 ± 0.2 ; *elav-Gal4/EP(3)381^{faf}*, 6.6 ± 0.2 ; *elav-Gal4/EP(3)381*, 9.4 ± 0.4 ; *elav-Gal4/EP(3)3520*, 7.9 ± 0.4 ; $P < 0.01$, $n = 54, 57, 48$ and 47 , respectively). Postsynaptic expression of *faf* does not affect synaptic morphology (data not shown).

To assess the physiological consequence of neuronal *faf* overexpression, we analysed both spontaneous and evoked neurotransmitter release at muscle 6 of third instar larvae (Fig. 2). Despite the greatly expanded size of the NMJ with *faf* overexpression, the amplitude of evoked excitatory junctional potentials (EJPs) was markedly reduced (*elav-Gal4/+*, 10.4 ± 1.7 mV; *elav-Gal4/EP(3)381*, 2.2 ± 0.3 mV; *elav-Gal4/EP(3)3520*, 2.3 ± 0.4 mV; $P < 0.001$, $n = 11, 10$ and 11 , respectively). Given that the amplitude of miniature EJPs (mEJPs) showed only a small, albeit significant, reduction (*elav-Gal4/+*, 0.90 ± 0.06 mV; *elav-Gal4/EP(3)381*, 0.73 ± 0.04 mV; *elav-Gal4/EP(3)3520*, 0.67 ± 0.05 mV; $P < 0.05$,

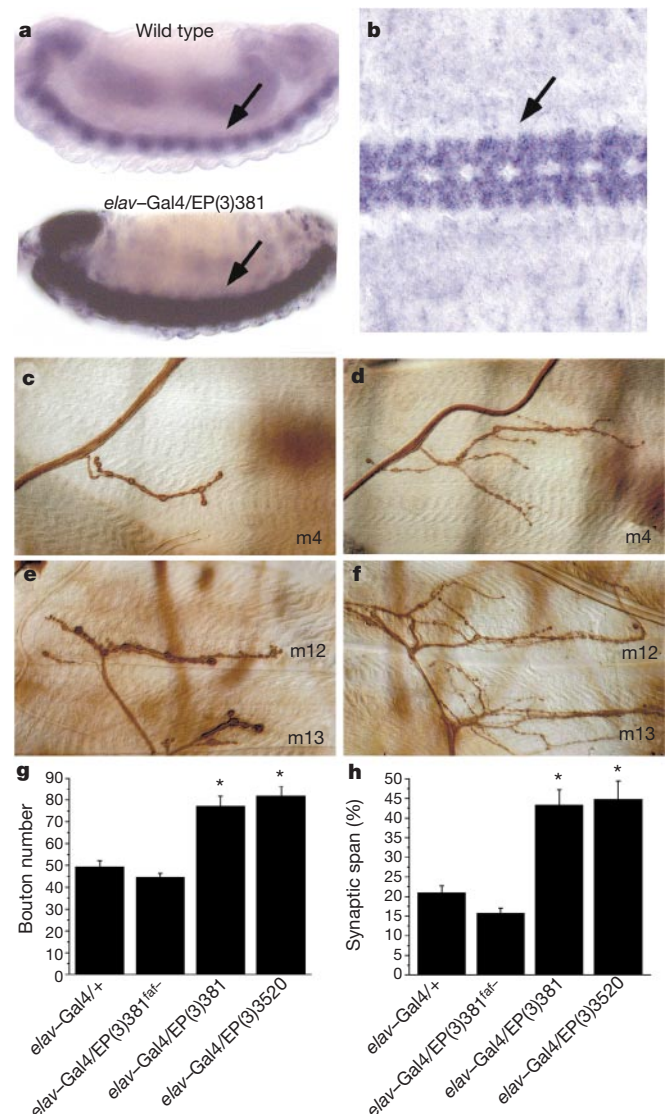


Figure 1 Neuronal overexpression of *fat facets* leads to synaptic overgrowth. **a, b**, RNA *in situ* hybridization shows that *faf* transcript (**a**) is overexpressed when EP(3)381 is crossed to *elav-Gal4*, and that endogenous *faf* transcript (**b**) is highly enriched in the developing CNS. Arrows indicate the developing CNS. **c–f**, Anatomical examination of NMJs reveals a marked expansion in the size and complexity of the synapse on neuronal overexpression of *faf* (*elav-Gal4/+* (**c, e**); *elav-Gal4/EP(3)381* (**d, f**)).

g, h, There is a significant increase in both the number of synaptic boutons (**g**) and the synaptic span (**h**) in animals overexpressing *faf* from either of two, independent EP lines (EP(3)381/*elav-Gal4* or EP(3)3520/*elav-Gal4*) compared with animals that do not overexpress *faf* (*elav-Gal4/+*) or that overexpress a *faf* transcript encoding a non-functional protein (EP(3)381^{faf}/*elav-Gal4*). The statistical test used was the *t*-test. Bouton number, $n = 27, 33, 32$ and 34 , respectively; synaptic span, $n = 18, 20, 16$ and 16 , respectively. *, $P < 0.001$; data represent the mean $\pm$ s.e.m.

$n = 11, 10$ and 11 , respectively), we measured a large decrease in quantal content (the number of vesicles released by the nerve) as measured by dividing the EJP amplitude by the mEJP amplitude (Fig. 2b). Neuronal overexpression of *faf* also leads to a reduction in the frequency of spontaneous mEJPs (Fig. 2b). The reduction in both quantal content and mEJP frequency indicates a presynaptic defect in neurotransmitter release. Other presynaptic mutants with even greater reductions in quantal content do not show a structural overgrowth⁸, thus the anatomical phenotype described above is probably a direct consequence of *faf* overexpression and not a secondary consequence of this physiological phenotype.

faf antagonizes ubiquitin-dependent mechanisms by deubiquitinating target proteins^{4,9}. Alterations in synaptic structure and function owing to overexpression of *faf* suggest that ubiquitin-dependent mechanisms normally act to regulate the developing synapse. However, *faf*, a characterized deubiquitinating protease, might have other functions. To investigate the role of deubiquitination in the regulation of synaptic development, we generated transgenic flies capable of the targeted overexpression of the yeast deubiquitinating protease UBP2 (ref. 5). This enzyme antagonizes ubiquitin-dependent mechanisms in yeast and has overlapping substrate specificity with FAF⁹. Overexpression of yeast UBP2 in the nervous system of *Drosophila* leads to marked synaptic overgrowth and a severe reduction in presynaptic transmitter release (Fig. 3a, b) (quantal content: *elav-Gal4/+*, 15.4 ± 1.3 ; *elav-Gal4/pUAS-UBP2*, 4.4 ± 1.1 ; $P \ll 0.001$; mEJP frequency (Hz): *elav-Gal4/+*, 2.3 ± 0.2 ; *elav-Gal4/pUAS-UBP2*, 1.2 ± 0.1 ; $P \ll 0.001$; EJP amplitude (mV): *elav-Gal4/+*, 17.5 ± 2.7 ; *elav-Gal4/pUAS-UBP2*, 3.8 ± 0.8 ; $P \ll 0.001$; mEJP amplitude (mV): *elav-Gal4/+*, 1.08 ± 0.09 ; *elav-Gal4/pUAS-UBP2*, 0.84 ± 0.06 ; $P < 0.05$, $n = 8$ and 10 , respectively). This phenotype is very similar to that seen with *faf* overexpression. Hence, antagonizing ubiquitin-dependent mechanisms by overexpression of deubiquitinating proteases markedly affects synaptic development.

To identify molecular pathways regulated by *faf*, we performed a genetic interaction screen to identify genes that enhance the *faf* overexpression phenotype. We screened the X chromosome for viable mutations that were lethal in combination with neuronal overexpression of *faf*. We screened 7,000 chromosomes and iden-

tified 15 lethal enhancers, 12 of which form one complementation group. These 12 mutants are alleles of the *highwire* (*hiw*) gene⁶ and share the synaptic overgrowth phenotype described for loss-of-function *hiw* mutants (data not shown). The *hiw* loss-of-function phenotypes are very similar to the *faf* gain-of-function phenotypes described here, with a large increase in the number of synaptic boutons, branches, and synaptic span, a small decrease in quantal size, and a large decrease in quantal content⁶. The *hiw* transcript encodes a greater than 5,000 amino-acid protein that is localized to synapses and that contains a RING-H2 finger⁶, a domain recently identified in a large family of E3 ubiquitin ligases¹⁰. Hence, a potential synaptic E3 ligase was identified as a lethal enhancer of neuronal overexpression of *faf*. This genetic interaction provides further evidence that ubiquitination may have a central role in regulating synaptic growth and function.

To further investigate the genetic relationship between *hiw* and *faf*, we generated double mutants between loss-of-function alleles of *hiw* and *faf*. *faf* loss-of-function mutants have phenotypes in the developing eye and female germ line¹¹. In *faf* mutants we found no defects in either synaptic morphology or function (data not shown), possibly due to genetic redundancy between *faf* and one of the 17 other putative deubiquitinating proteases in the *Drosophila* genome¹². Although we found no phenotype for *faf* mutants in otherwise wild-type flies, in the sensitized background of a *hiw* mutant we do find a requirement for *faf*. Two different loss-of-function alleles of *faf* both suppress the physiological phenotype of *hiw*, leading to a more than doubling of both quantal content and mEJP frequency (Fig. 4) (EJP amplitude (mV): wild type, 30.7 ± 1.0 ; *hiw*, 7.2 ± 0.7 ; *hiw; faf^{BX4}/faf^{BX4}*, 16.7 ± 0.6 ; *hiw; faf^{BX4}/faf^{F08}*, 13.9 ± 0.7 ; mEJP amplitude (mV): wild type, 0.96 ± 0.04 ; *hiw*, 0.64 ± 0.03 ; *hiw; faf^{BX4}/faf^{BX4}*, 0.76 ± 0.04 ; *hiw; faf^{BX4}/faf^{F08}*, 0.66 ± 0.04 ; $n = 18, 16, 18$ and 14 , respectively). Hence endogenous *faf* activity acts to inhibit neurotransmitter release in a *hiw* background, much as increased *faf* activity inhibits neurotransmitter release in an otherwise wild-type background (Fig. 2). Mutants of *faf* do not suppress the synaptic overgrowth seen in *hiw*, indicating that the physiological and morphological phenotypes in *hiw* are genetically separable. This suggests that either these phenotypes are mediated by different *hiw* substrates, or the anatomical phenotype is

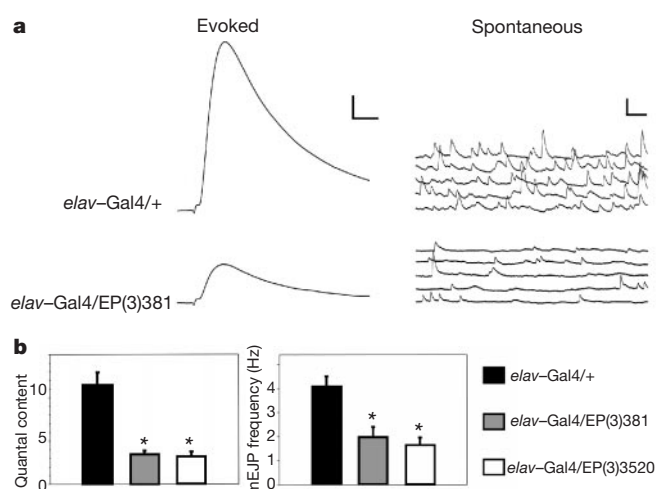


Figure 2 Neuronal overexpression of *fat facets* impairs synaptic transmission. **a**, Representative traces of evoked EJPs and spontaneous mEJPs from the indicated genotypes. **b**, There is a significant decrease in both the quantal content and mEJP frequency in genotypes overexpressing *faf* (EP(3)381/*elav-Gal4* and EP(3)3520/*elav-Gal4*) compared with controls (*elav-Gal4/+*) (*t*-test, $*$, $P < 0.001$, $n = 11, 10$ and 11 , respectively). Data represent the mean $\pm$ s.e.m. Calibration: 1.5 mV/15 ms evoked; 2 mV/150 ms spontaneous.

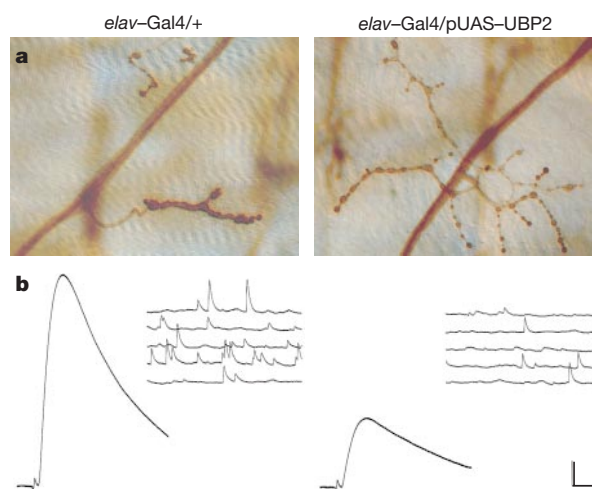


Figure 3 Neuronal overexpression of the yeast deubiquitinating protease UBP2 phenocopies neuronal overexpression of *faf*. **a**, Anatomical examination of NMJs reveals an expansion in the size and complexity of the synapse when the yeast deubiquitinating protease UBP2 is expressed in the *Drosophila* CNS. **b**, Representative traces of evoked and spontaneous EJPs and mEJPs recorded from the indicated genotypes. Expression of UBP2 leads to a significant decrease in both the calculated quantal content and the frequency of spontaneous mEJPs. Calibration: 2 mV/15 ms evoked; 1.5 mV/275 ms spontaneous.

more sensitive to disruption of ubiquitin-dependent mechanisms. Finally, the inability to suppress the small quantal size defect suggests that this phenotype, seen in both *hiw* mutants and with overexpression of *faf*, may be a secondary consequence of synaptic overgrowth.

Decreases in postsynaptic activity induce a compensatory increase in presynaptic transmitter release, demonstrating that a homeostatic mechanism regulates synaptic strength during development^{13,14}. To assess the relationship between homeostatic and ubiquitin-dependent regulation, we disrupted glutamate receptor function in a *hiw* mutant. Postsynaptic expression of a dominant negative glutamate receptor (*DgluRIIA-M/R*)¹⁴ in a *hiw* mutant leads to an 18% decrease in quantal size (*hiw*, 0.65 ± 0.03 mV; *hiw*; *DgluRIIA-M/R*, 0.53 ± 0.03 mV; $P < 0.01$), no change in the amplitude of the evoked response (*hiw*, 7.2 ± 0.7 mV; *hiw*; *DgluRIIA-M/R*, 7.4 ± 0.6 mV), and a compensatory 30% increase in quantal content (*hiw*, 10.8 ± 0.8 mV; *hiw*; *DgluRIIA-M/R*, 14.0 ± 0.9 mV; $P < 0.02$) (t -test; $n = 16$ and 10 , respectively). As homeostatic compensation still occurs in a *hiw* mutant, we suggest that homeostatic and ubiquitin-dependent regulation are mechanistically distinct.

The data presented here indicate that ubiquitin-dependent mechanisms regulate synaptic development at the *Drosophila* NMJ and suggest that a balance between positive and negative regulators of ubiquitination controls the structure and function of the synapse. Antagonizing ubiquitination by the neuronal overexpression of the deubiquitinating proteases *faf* or yeast UBP2 leads to synaptic overgrowth and defects in neurotransmitter release. This phenotype is very similar to the loss-of-function phenotype of *hiw*, a putative synaptic E3 ubiquitin ligase. Gain-of-function mutants of *faf* enhance *hiw* and loss-of-function alleles of *faf* suppress *hiw*. We propose that *hiw*-dependent ubiquitination controls the level or activity of critical regulatory molecules at the synapse, and that these molecules can be deubiquitinated by *faf* and other deubiquitinating proteases. Ubiquitinated proteins have been identified at mammalian synapses^{15,16}, and ubiquitin-processing enzymes can regulate long-term potentiation and facilitation^{17,18}, therefore control of ubiquitination by molecules such as HIW and FAF could be a widely used mechanism for regulating synaptic growth and function. □

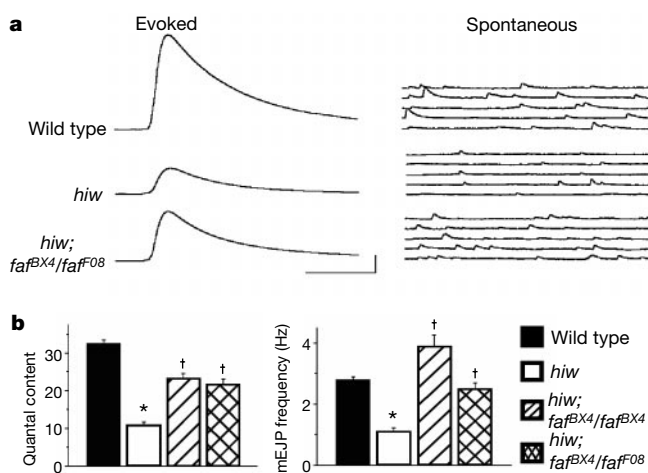


Figure 4 Loss-of-function mutations in *fat facets* suppress *highwire*. **a**, Representative traces of evoked EJPs and spontaneous mEJPs from the indicated genotypes. **b**, There is a significant reduction in the quantal content and mEJP frequency in *hiw* mutants compared with wild type (*, t -test, $P < 0.001$). This deficit in *hiw* is rescued in *hiw*; *faf* double mutants; there is a greater than 100% increase in both quantal content and mEJP frequency (†, t -test, $P < 0.001$; $n = 18, 16, 18$ and 14 , respectively). Data represent the mean $\pm$ s.e.m. Calibration: 5 mV/20 ms evoked; 5 mV/200 ms spontaneous.

Methods

Molecular genetics

We identified EP(3)381 and EP(3)3520 lines in a screen for synaptic morphology mutants. The Rorth collection⁷ of 2293 EP lines was crossed to the pan-neuronal¹⁹ driver *elav-Gal4*, and lines with behavioural phenotypes were identified. From this group, a secondary screen was performed to assess NMJ structure. Both EP(3)381 and EP(3)3520 (originally identified owing to an abnormal gait) produce expanded synapses when crossed to *elav-Gal4*. Both lines are located upstream of the *faf* gene and are oriented to overexpress the gene. We confirmed this overexpression with *in situ* hybridization. EP(3)381^{faf} was generated by the EP(3)381 chromosome undergoing mutagenesis with ethylmethane sulphonate (EMS), crossing to the *faf* mutant *faf*^{F08}, and identifying F₁ progeny with rough eyes. EP(3)381^{faf} fails to complement both the eye and female sterility phenotypes of two *faf* alleles (*faf*^{BX4} and *faf*^{F08}), demonstrating that EP(3)381^{faf} is a new allele of *faf*. *In situ* hybridization demonstrates that EP(3)381^{faf} is still capable of overexpressing *faf* transcript. *highwire* was identified as a recessive enhancer of neuronal *faf* overexpression. In brief, *w*-males underwent mutagenesis with EMS, were crossed to *run/FM7TW9*, and 7,000 **FM7TW9* females were collected (* indicates a unique but unidentified mutation). These were mated singly to *w*-*Y*; *elav-Gal4*(*w*+), EP(3)381(*w*+)/TM6B males. Lines were identified in which **Y* was lethal in combination with *elav-Gal4*(*w*+), EP(3)381(*w*+), and viable with TM6B. Twelve such enhancers fail to complement the original *hiw* allele. We made the pUAS-UBP2 transgene by inserting the UBP2 gene⁹ into the pUAST vector²⁰.

Immunocytochemistry

Analysis of bouton number, synaptic span and branch point number was performed after anti-synaptotagmin (gift of K. Zinn) and anti-fasciclin II staining as described²¹. Bouton number was measured at muscle 4 in segment A3, whereas synaptic span was measured at muscle 13 and branch point number was quantified at muscle 12. No significant differences in muscle size were observed for any genotype and so no corrections were made for muscle size. Photographs of NMJ morphology were taken from larvae stained with anti-Syt and anti-Hrp (1:1000; Cappel).

Electrophysiology

Intracellular recordings were made from muscle 6, segment A3 as described¹³. Larvae were dissected in Stewart saline HL3 (ref. 22) containing 0.4 mM Ca²⁺ (Fig. 2), 0.45 mM Ca²⁺ (Fig. 3) and 0.6 mM Ca²⁺ (Fig. 4). In Fig. 4 all genotypes expressed a green fluorescent protein (GFP)-tagged marker of synaptic morphology²³. Cells with an input resistance of greater than 5 M Ω and a resting potential of less than -60 mV were selected for further analysis. No significant differences were observed for input resistance or membrane potential. We analysed greater than 30 EJPs and 100 s of mEJP recordings (pClamp8, Axon Instruments; Jaesin Software). The mean quantal content was estimated by dividing the mean EJP amplitude by the mean mEJP amplitude for each cell. No corrections for non-linear summation were made because such a correction would only exaggerate the reported phenotypes.

Received 17 April; accepted 31 May 2001.

- Hochstrasser, M. Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* **30**, 405–439 (1996).
- Wilkinson, K. D. Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB J.* **11**, 1245–1256 (1997).
- Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
- Huang, Y., Baker, R. T. & Fischer-Vize, J. A. Control of cell fate by a deubiquitinating enzyme encoded by the fat facets gene. *Science* **270**, 1828–1831 (1995).
- Baker, R. T., Tobias, J. W. & Varshavsky, A. Ubiquitin-specific proteases of *Saccharomyces cerevisiae*. Cloning of UBP2 and UBP3, and functional analysis of the UBP gene family. *J. Biol. Chem.* **267**, 23364–23375 (1992).
- Wan, H. I. et al. *Highwire* regulates synaptic growth in *Drosophila*. *Neuron* **26**, 313–329 (2000).
- Rorth, P. et al. Systematic gain-of-function genetics in *Drosophila*. *Development* **125**, 1049–1057 (1998).
- DiAntonio, A. & Schwarz, T. L. The effect on synaptic physiology of synaptotagmin mutants in *Drosophila*. *Neuron* **12**, 909–920 (1994).
- Wu, Z., Li, Q., Fortini, M. E. & Fischer, J. A. Genetic analysis of the role of the *Drosophila* fat facets gene in the ubiquitin pathway. *Dev. Genet.* **25**, 312–320 (1999).
- Joazeiro, C. A. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* **102**, 549–552 (2000).
- Fischer-Vize, J. A., Rubin, G. M. & Lehmann, R. The fat facets gene is required for *Drosophila* eye and embryo development. *Development* **116**, 985–1000 (1992).
- Chen, X. & Fischer, J. A. *In vivo* structure/function analysis of the *Drosophila* fat facets deubiquitinating enzyme gene. *Genetics* **156**, 1829–1836 (2000).
- Petersen, S. A., Fetter, R. D., Noordermeer, J. N., Goodman, C. S. & DiAntonio, A. Genetic analysis of glutamate receptors in *Drosophila* reveals a retrograde signal regulating presynaptic transmitter release. *Neuron* **19**, 1237–1248 (1997).
- DiAntonio, A., Petersen, S. P., Heckmann, M. & Goodman, C. S. Glutamate receptor expression regulates quantal size and quantal content at the *Drosophila* neuromuscular junction. *J. Neurosci.* **19**, 3023–3032 (1999).
- Serdaroglu, P., Askanas, V. & Engel, W. K. Immunocytochemical localization of ubiquitin at human neuromuscular junctions. *Neuropathol. Appl. Neurobiol.* **18**, 232–236 (1992).
- Chapman, A. P., Smith, S. J., Rider, C. C. & Beesley, P. W. Multiple ubiquitin conjugates are present in rat brain synaptic membranes and postsynaptic densities. *Neurosci. Lett.* **168**, 238–242 (1994).
- Hegde, A. N. et al. Ubiquitin C-terminal hydrolase is an immediate-early gene essential for long-term facilitation in Aplysia. *Cell* **89**, 115–126 (1997).
- Jiang, Y. H. et al. Mutation of the Angelman ubiquitin ligase in mice causes increased cytoplasmic p53 and deficits of contextual learning and long-term potentiation. *Neuron* **21**, 799–811 (1998).

19. Yao, K. M. & White, K. Neural specificity of elav expression: defining a *Drosophila* promoter for directing expression to the nervous system. *J. Neurochem.* **63**, 41–51 (1994).
20. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
21. Schuster, C. M., Davis, G. W., Fetter, R. D. & Goodman, C. S. Genetic dissection of structural and functional components of synaptic plasticity. I. Fasciclin II controls synaptic stabilization and growth. *Neuron* **17**, 641–654 (1996).
22. Stewart, B. A., Atwood, H. L., Renger, J. J., Wang, J. & Wu, C. F. Improved stability of *Drosophila* larval neuromuscular preparations in haemolymph-like physiological solutions. *J. Comp. Physiol. A* **175**, 179–191 (1994).
23. Zito, K., Parnas, D., Fetter, R. D., Isacoff, E. Y. & Goodman, C. S. Watching a synapse grow: noninvasive confocal imaging of synaptic growth in *Drosophila*. *Neuron* **22**, 719–729 (1999).

Acknowledgements

We would like to thank J. Fischer and K. Zinn for their gifts of reagents. This work was supported by the NIH and HHMI (C.S.G.), NSERC Canada (A.P.H.), and the Burroughs-Wellcome Career Award, HHMI, and Whitehall Foundation (A.D.).

Correspondence and requests for material should be addressed to A.D. (e-mail: diananton@pcg.wustl.edu).

Antibacterial agents based on the cyclic D,L- α -peptide architecture

Sara Fernandez-Lopez, Hui-Sun Kim, Ellen C. Choi, Mercedes Delgado, Juan R. Granja*, Alisher Khasanov, Karin Kraehenbuehl, Georgina Long, Dana A. Weinberger, Keith M. Wilcoxon & M. Reza Ghadiri

Departments of Chemistry and Molecular Biology and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California 92037, USA

The rapid emergence of bacterial infections that are resistant to many drugs underscores the need for new therapeutic agents^{1–3}. Here we report that six- and eight-residue cyclic D,L- α -peptides act preferentially on Gram-positive and/or Gram-negative bacterial membranes compared to mammalian cells, increase membrane permeability, collapse transmembrane ion potentials, and cause rapid cell death. The effectiveness of this class of materials as selective antibacterial agents is highlighted by the high efficacy observed against lethal methicillin-resistant *Staphylococcus aureus* infections in mice. Cyclic D,L- α -peptides are proteolytically stable, easy to synthesize, and can be derived from a potentially vast membrane-active sequence space. The unique abiotic structure of the cyclic peptides and their quick bactericidal action may also contribute to limit temporal acquirement of drug resistant bacteria. The low molecular weight D,L- α -peptides offer an attractive complement to the current arsenal of naturally derived antibiotics, and hold considerable potential in combating a variety of existing and emerging infectious diseases.

Cyclic D,L- α -peptides^{4,5} possess unique structural features not found in the natural class of peptide antibiotics and/or their derivatives^{6–9}. Cyclic peptides with an even number of alternating D- and L- α -amino acids can adopt flat, ring-shaped conformations in which the backbone amide functionalities are oriented perpendicular to the side chains and the plane of the ring structure (Fig. 1). Under conditions that favour hydrogen bonding, such as adsorption onto lipid membranes^{10–12}, the cyclic peptides can stack to form hollow, β -sheet-like tubular structures that are open-ended, presenting the amino acid side chains on the outside surface of the ensemble. Therefore, we proposed that appropriately designed cyclic D,L- α -peptides may be able to selectively target and self

assemble in bacterial membranes and exert antibacterial activity by increasing the membrane permeability¹³. The studies reported here are consistent with this hypothesis.

We observed that amphiphilic cyclic D,L- α -peptides composed of three consecutive hydrophilic residues and repeating L-tryptophan and D-leucine amino acids self assemble in synthetic membranes into tubular structures allowing for highly efficient transport of ions and small species with relative molecular masses of less than 10,000 (M_w 10K) across lipid bilayers—as demonstrated by vesicular ion and molecular transport assays, conductance measurements in planar lipid bilayers, and Fourier transform–infrared spectroscopy (FT–IR) analyses¹³. Peptide 4, a representative member of this class of cyclic peptides, displayed potent *in vitro* activity against Gram-positive *Bacillus subtilis* and *Staphylococcus aureus*, and against Gram-negative *Streptococcus pneumoniae* and vancomycin-resistant *Enterococcus faecalis* (Table 1). On the basis of these results, we designed a series of six- and eight-residue amphiphilic cyclic D,L- α -peptides to further probe the scope and use of this class of peptides as antibacterial agents, and to examine the relationship between surface properties, antibacterial activities and membrane selectivity (Table 2). Each peptide was designed to bear at least one basic residue to enhance its target specificity towards negatively charged bacterial membranes^{7,8}. Antibacterial activities were tested against *Escherichia coli* and methicillin-resistant *S. aureus* (MRSA)—currently the principal component of roughly two million annual nosocomial infections in the United States (<http://narsaweb.narsa.net>). We modelled selectivity for bacterial versus mammalian cells by comparing antibacterial activity to haemolytic activity in red blood cells¹⁴.

The role of hydrophilic side chains in biological activity and membrane selectivity is evident on examination of the antibacterial activities listed in Table 2. Cyclic peptides 1 and 2 each bear one

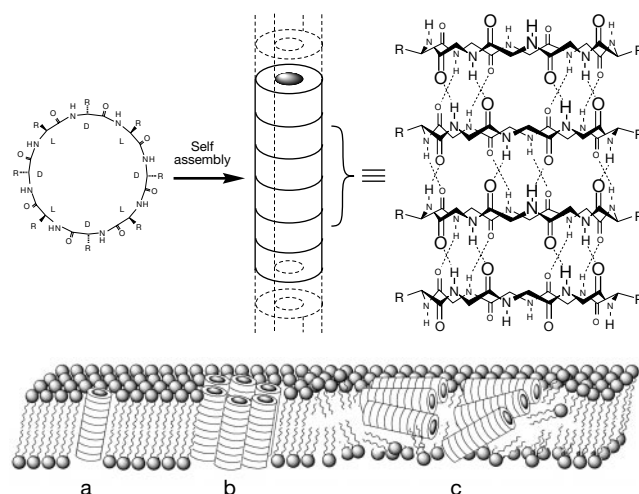


Figure 1 The β -sheet-like, hydrogen-bonded tubular architecture of the self-assembled, eight-residue cyclic D,L- α -peptide, and modes of membrane permeation accessible to peptide nanotubes. Top, schematic representation of an eight-residue cyclic D,L- α -peptide subunit. Note the flat, ring-shaped conformation with side chains (R) decorating the outside surface. Self-assembly is directed by intersubunit backbone–backbone hydrogen bonding to give β -sheet-like, tubular supramolecular structures that are open-ended and hollow. Most side chains are omitted for clarity. Bottom, cyclic D,L- α -peptide nanotubes can display sequence-dependent modes of membrane permeation: a, intramolecular pore; b, barrel stave; and c, carpet-like (cyclic peptides are depicted as ring structures). It is thought that peptide nanotubes operating through the carpet-like mode of action would have a greater potential for membrane discrimination because of its polyvalent display of surface-exposed hydrophilic side chains for potential interactions with various membrane constituents.

* Present address: Departamento de Química Orgánica, Universidad de Santiago de Compostela, 15706 Santiago de Compostela, Spain.

basic residue between two serines or between a serine and a threonine residue, respectively. Although peptide 1 displays good activity against MRSA, substitution of lysine for histidine in peptide 2 significantly decreases the activity. Similarly, cyclic peptides 4–11 each possess two basic and one neutral polar residue, and display about 15-fold variations in antibacterial activity and up to a fivefold difference in haemolytic activity. Glutamic acid substitution has a deleterious effect on the antibacterial activity, as indicated by the large difference in the activity between peptides 10 and 12, which differ only by a serine to glutamic acid mutation. The reduced activity of the peptide that contains glutamic acid is probably due to unfavourable electrostatic interactions of the carboxylate side chain with the bacterial membrane constituents. Increasing the number of basic residues from two to three in peptides 13–18 yields high activities against MRSA, with sequences 14 and 17 exhibiting moderate activities against *E. coli*. *In vitro* antibacterial activities of hexameric sequences 22–25 also reflect the relationship between the outside surface properties of peptide nanotubes and activity level, and membrane selectivity. Peptide 22, which has two consecutive lysine residues, displays broad-spectrum activity and low haemolytic propensity. On the other hand, peptide 23, which differs from 22 by a lysine to histidine substitution, retains high activity against MRSA but is inactive against *E. coli* and is highly haemolytic. Substituting a lysine in peptide 22 with serine yields the inactive sequence 24. Notably, peptide 25, which possesses two arginine residues, displays potent and selective activity against *E. coli*, with low levels of haemolysis. The broad spectrum of activity and membrane selectivity observed with the above peptides highlights the effects that single amino-acid substitutions have on membrane activity and selectivity.

Fluorescence-based cell depolarization studies, the activity of the cyclic peptides compared to that of analogous linear peptides, the rate of bacteria killing, and attenuated total reflectance (ATR) FT-IR studies in synthetic lipids are all consistent with a membrane permeation mode of action for this class of antibacterial peptides. The cyanine membrane dye 3,3'-dipropylthiadicarbocyanine iodide (diSC₃(5)), for which the fluorescence spectrum is sensitive to changes in membrane potential, was used to analyse the effects of peptides 4, 14 and 22 on live *S. aureus*^{15–17}. In each case, the exposure of dye-saturated live *S. aureus* to cyclic peptides at various concentrations (0.1–10× minimum inhibitory concentration (MIC)) leads to fast and complete membrane depolarization (Fig. 2). Moreover, culture samples taken during the course of these experiments correlate the membrane depolarization with cell death at peptide concentrations equal to or above the MIC. No living bacteria are

detectable after 5 min exposure to MIC levels of peptides, whereas at 0.1×MIC levels a significant population of live bacteria is still present.

Studies with control peptides 19–21—linear versions of the active sequence 13—show little *in vitro* activity against MRSA. Furthermore, the positively charged control peptide *cyclo*((L-Lys-D-Leu)₄), which is devoid of the amphiphilic sequence arrangement, is membrane inactive and lacks *in vitro* antibacterial activity. The inactivity of these control peptides supports the role of cyclic D,L-α-peptide structure and conformation in the formation of functional tubular-membrane assembly in bacteria. The possibility of a receptor/ligand-mediated mode of action for the cyclic peptides in bacterial membranes is unlikely as both enantiomeric sequences 4 and 5 have similar *in vitro* activities, and because the rates of bacteria killing are consistent with this mode of action^{18,19}. Specifically, time-killing studies with octameric peptides 4 and 15 and hexameric peptides 22 and 23 at various concentrations show that at concentrations equal to or above the MIC, the octameric and hexameric peptides have complete bactericidal activity against MRSA within 5 and 60 min, respectively. This timescale is more consistent with membrane permeation than a receptor/ligand-mediated binding/inhibition mechanism that typically takes several hours to achieve complete bactericidal or bacteriostatic activity. ATR/FT-IR spectroscopy indicates that cyclic peptides 4, 13 and 14 self assemble to form nanotubes in synthetic lipid membranes¹² (Table 3). The infrared spectra display amide-A (NH stretch), amide-I and amide-II bands that are characteristic of antiparallel, β-sheet nanotubular structures that are tightly hydrogen bonded¹². Quantitative measurements in oriented dimyristoyl phosphatidylcholine (DMPC) lipid multi-bilayers indicate that peptide nanotubes are

Table 1 *In vitro* activity of peptide 4 (KQRWLWLW)

Bacterial/cell type	MIC (μg ml ⁻¹)
Gram-positive bacteria*	
<i>Bacillus subtilis</i>	3†
<i>Bacillus cereus</i>	2
<i>Staphylococcus aureus</i>	3†
<i>Listeria monocytogenes</i>	20
Gram-negative bacteria†	
<i>Enterococcus faecalis</i>	10†
<i>Streptococcus pneumoniae</i>	10†
<i>Salmonella typhimurium</i>	>70
Mammalian cells	
Red blood cell (mouse)	45§
M21 (melanoma)	>75

Underlined amino acids indicate D-amino-acid residues. MIC, minimum inhibitory concentration. * *Bacillus subtilis* (ATCC 43223); *B. cereus* (ATCC 11778); *S. aureus* (ATCC 25923); *L. monocytogenes* (ATCC 19115).

† Data from MDS Panlabs Pharmacology Services.

‡ *Enterococcus faecalis* (vancomycin-resistant clinical isolate); *S. pneumoniae* (erythromycin- and ampicillin-resistant clinical isolate); *S. typhimurium* (ATCC 14028).

§ HD₅₀ (50% haemolytic dose, μg ml⁻¹).

|| LD₅₀ (50% lethal dose, μg ml⁻¹).

Table 2 *In vitro* activity of D,L-α-amino acid cyclic peptides

Peptides	<i>S. aureus</i> (MRSA)* MIC (μg ml ⁻¹)	<i>E. coli</i> † MIC (μg ml ⁻¹)	Haemolysis‡ HD ₅₀ (μg ml ⁻¹)
1 (SKSWLWLW)	10	100	30
2 (THSWLWLW)	100	100	100
3 (RGDWLWLW)	100	100	100
4 (KQRWLWLW)	6	80	45
5 (KQRWLWLW)	8	80	40
6 (RQRWLWLW)	18	90	25
7 (KQKWLWLW)	45	100	50
8 (KSKWLWLW)	5	40	100
9 (SHKWLWLW)	10	100	35
10 (SKHWLWLW)	15	100	20
11 (SHHWLWLW)	20	100	20
12 (EKHWLWLW)	100	100	>100
13 (KKKWLWLW)	8	70	50
14 (RRKWLWLW)	6	15	50
15 (KRKWLWLW)	10	40	50
16 (RRRWLWLW)	10	50	35
17 (HKHWLWLW)	12	15	25
18 (KHKWLWLW)	10	80	30
19 H ₂ N-VK ⁺ KKWLWLW-COOH	60	–	100
20 H ₂ N-KKKWLWLW-CONH ₂	50	–	100
21 Ac-KK ⁺ KKWLWLW-CONH ₂	80	–	100
22 (KKLWLW)	10	17	80
23 (KH ⁺ LWLW)	10	100	25
24 (KSLWLW)	75	100	90
25 (RRLWLW)	35	5	90

Single-letter codes for amino acids and shorthand representation of cyclic peptide sequences are used to allow comparisons between peptide sequences. The brackets indicate cyclic structure and underlining represents D-amino-acid residues. Data for the most active and selective sequences are in bold. Peptides were synthesized by standard solid-phase BOC (t-butoxycarbonyl) or Fmoc (fluorenylmethoxycarbonyl) synthetic protocols, cyclized in solution or on solid support, purified by reverse phase HPLC, and characterized by matrix-assisted laser desorption/ionization or electrospray ionization mass spectrometry.

* ATCC 33591.

† JM109 (DE3).

‡ See Methods for haemolysis assay condition. Antibacterial assays were performed according to standard microdilution methods²².

oriented at a $70 \pm 5^\circ$ tilt angle from the membrane normal, consistent with the carpet-like mode of action^{7,13} (Fig. 1).

We next examined the effect of plasma proteins on the availability and stability of cyclic peptides *in vitro*. Despite the hydrophobicity of the peptides studied, antibacterial activities remained unchanged in the presence of large amounts (up to 50% v/v) of fetal bovine serum (FBS) in the culture media. Moreover, most peptides examined have a reduced level of haemolytic activity in the presence of 5–10% FBS in the assay mixture, compared with analogous assays in the absence of FBS. For example, peptide 4 in the presence of 10% FBS displays a fivefold reduction in haemolytic activity (from $HD_{50} = 10$ to $50 \mu\text{g ml}^{-1}$; HD_{50} is the 50%-haemolytic dose). Furthermore, cyclic peptides 9, 14, 22 and 25 were assayed for proteolytic susceptibility. As predicted from their abiotic structure and conformational preferences, these peptides were stable in the presence of trypsin, α -chymotrypsin, subtilisin and murine blood plasma. No significant peptide degradation was observed in chromatograms obtained from reverse-phase high-performance liquid chromatography (HPLC) over a 72-h time period, whereas control linear L- α -amino-acid peptides were degraded in less than 10 min under similar reaction conditions (4 h for murine blood plasma).

Limited toxicology studies in mice were conducted to evaluate various routes of drug administration, maximum tolerable dose, and blood and tissue toxicity. Two mice in each dose study and two control animals for each study were injected with a bolus intravenous, intraperitoneal, or subcutaneous dose of peptide 4, 14 or 22. Studies with peptide 4 indicate that bolus intravenous injections of 12 mg per kg cause signs of temporary (less than 10 min in duration) discomfort. In contrast, no signs of toxicity were observed at the maximum tested dose (12 mg per kg) with the intraperitoneal route of administration. Peptide 22 was tolerated up to the maximum tested intraperitoneal dose (50 mg per kg), with the highest dose causing immediate signs of acute toxicity that dissipated within 2 h. Peptide 14 was tested up to 17.5 mg per kg intraperitoneal and subcutaneous doses without any apparent signs of toxicity. A regimen of 300 mg per day of peptide 14 for three consecutive days was administered to two mice intraperitoneally and two mice subcutaneously. Over four days, no apparent changes in the physical, social and feeding activities were observed in these mice relative to control mice injected with vehicle without peptide. Moreover, haematology, necropsy and detailed microscopic examination of various tissue and organ samples carried out on the fourth day after the initial injection showed normal blood and morphological profiles, except for the sites of intraperitoneal and subcutaneous injections (K. G. Osborn, unpublished observations). These sites exhibited moderate subacute inflammation typical of intraperitoneal and subcutaneous drug administration.

A pilot study was conducted to determine the efficacy of peptide 14 with respect to *in vivo* protection of mice from bacterial infection. Three groups of mice (six mice per group) were infected intraperitoneally (left side) with a lethal dose of MRSA ($2-5 \times 10^7$

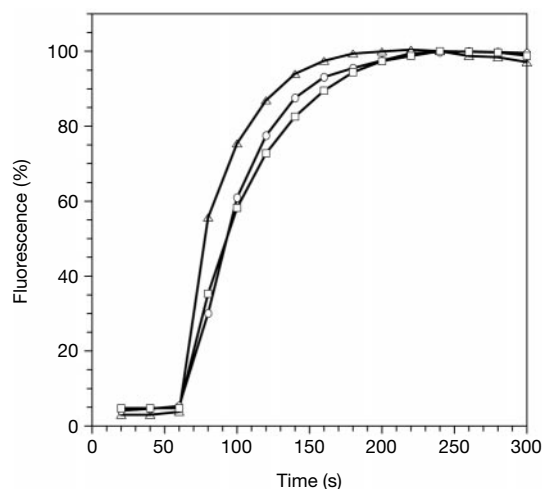


Figure 2 Rates of membrane depolarization of intact *S. aureus* (ATCC 25923) on exposure to self-assembling peptides. Membrane depolarization was monitored by the change in the intensity of fluorescence emission of the membrane-potential-sensitive dye diSC₃(5) (excitation wavelength $\lambda_{\text{ex}} = 622$ nm, emission wavelength $\lambda_{\text{em}} = 670$ nm on addition of minimum inhibitory concentrations of peptide 4 (circles), peptide 14 (triangles) and peptide 22 (squares). Peptides were added at 1 min and aliquots taken at 4 min to obtain viable bacterial counts. In all cases at least a 10^3 fold reduction in viability on the addition of the peptides was observed.

colony-forming units (CFU) per mouse) and then were administered a 13 mg per kg bolus intraperitoneal (right side) or subcutaneous dose of the peptide 45 min after MRSA injection. All mice in the control group receiving the vehicle without peptide died within 48 h, whereas 67% and 50% of mice that received the peptide intraperitoneally and subcutaneously, respectively, survived during the course of a seven day study. After this preliminary demonstration of *in vivo* antibacterial efficacy, we undertook a larger efficacy study with peptides 13, 14 and 22. Groups of mice were infected with lethal doses of MRSA (intraperitoneal, left side) and then each group was treated intraperitoneally with a given bolus dose of peptide 13, 14 or 22 at 45–60 min after initial infection. We observed the mice for seven days. In each case, a single appropriate dose of peptide was sufficient to completely protect various groups of mice from MRSA infections (Fig. 3). Notably, the effective doses parallel the *in vitro* bactericidal activities. Specifically, the PD_{50} (50%-protective dose) values of peptides 13, 14 and 22 are 6 ± 2 mg per kg, 5 ± 2 mg per kg and 10 ± 2 mg per kg, respectively. These *in vivo* studies highlight the considerable potential that this class of peptides of low molecular weight holds for the treatment of methicillin-resistant *S. aureus* infections.

The ability of cyclic D,L- α -peptides and the related β -amino-acid cyclic peptides^{20,21} to form a wide array of structurally analogous, biologically active structures with surface properties that are

Table 3 ATR/FT-IR data and orientation of antibacterial cyclic peptide nanotubes in oriented DMPC lipid multibilayers

Cyclic peptide	Frequency (cm^{-1})			Peptide		Lipid		$\Delta\text{Tilt}^\S$
	Amide-I	Amide-II	Amide-A	$[A_{\parallel}/A_{\perp}]^*$	Tilt ‡	$[A_{\parallel}/A_{\perp}]^\ddagger$	Tilt ‡	
4 (<u>K</u> <u>Q</u> <u>R</u> <u>W</u> <u>L</u> <u>W</u> <u>L</u> <u>W</u>)	1,629	1,538	3,283	1.16	70°	1.16	29°	41°
13 (<u>K</u> <u>K</u> <u>K</u> <u>W</u> <u>L</u> <u>W</u> <u>L</u> <u>W</u>)	1,629	1,537	3,281	1.13	71°	1.17	29°	42°
14 (<u>R</u> <u>R</u> <u>K</u> <u>W</u> <u>L</u> <u>W</u> <u>L</u> <u>W</u>)	1,628	1,538	3,280	1.34	65°	1.15	29°	36°

Underlined amino acids indicate D-amino-acid residues. DMPC, dimyristoyl phosphatidylcholine.

* Dichroic ratio of the amide-I intensity with parallel, polarized incident light to the band intensity with the perpendicular polarized light.

† Tilt refers to the angle of the molecular axis with respect to the surface normal.

‡ Dichroic ratio of the antisymmetric CH_2 stretches of DMPC lipid.

§ Difference between the angles of the peptide tube axis and the lipid hydrocarbon chain. Lipid and peptide nanotube tilt angles are calculated according to methods detailed in ref. 12. Data are average of two samples with errors less than $\pm 2^\circ$.

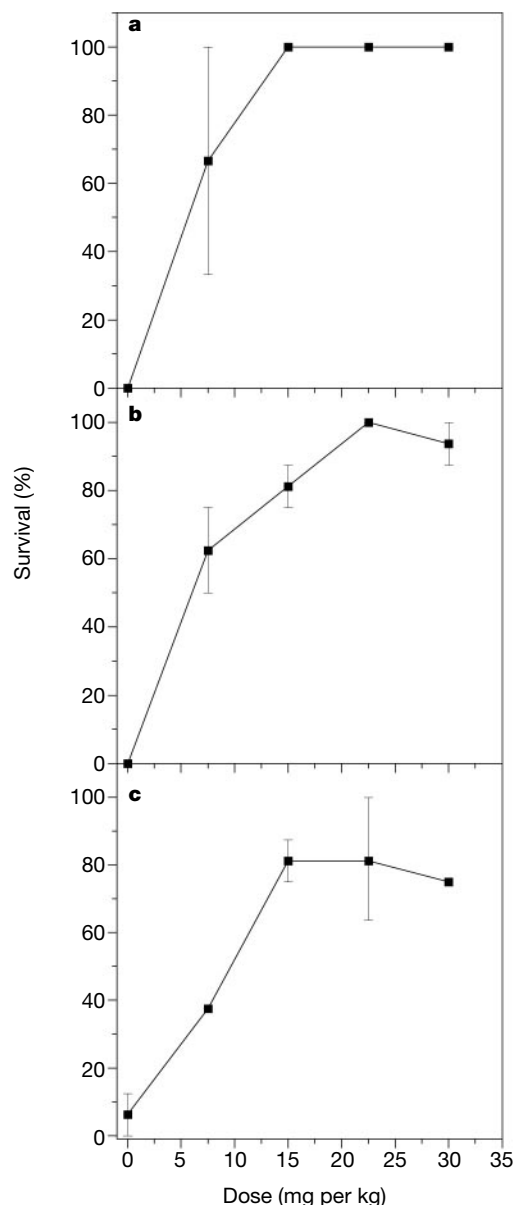


Figure 3 *In vivo* antibacterial efficacy of peptide nanotubes determined by survival rates of treated versus untreated mice challenged with lethal doses of methicillin-resistant *S. aureus*. **a**, peptide 14 ($PD_{50} = 5 \pm 2$ mg per kg). **b**, Peptide 13 ($PD_{50} = 6 \pm 2$ mg per kg). **c**, Peptide 22 ($PD_{50} = 10 \pm 2$ mg per kg). Values are mean $\pm$ s.d., $n = 2$.

tailored individually, coupled with the ease of the peptide synthesis, is expected to open new opportunities in the treatment of existing and emerging infectious diseases through both rational design and selection from combinatorial cyclic peptide libraries. □

Methods

Membrane depolarization assay

Bacteria were grown at 37 °C with agitation to mid log phase (absorbance at 600 nm (A_{600}) = 0.5). Cells were centrifuged and washed once with buffer (20 mM glucose, 5 mM HEPES pH 7.3) and resuspended to an A_{600} of 0.05 in a similar buffer containing 0.1 M KCl. Cells were incubated with 1 μ M diSC₃(5) until a stable reduction of fluorescence was achieved (around 15 min), indicating incorporation of the dye into the bacterial membrane. Peptides were added from stock solutions (1 mg ml⁻¹) dissolved in 5% DMSO in aqueous sucrose (9%) to achieve desired 0.1–10 $\times$ MIC concentrations.

In vitro haemolysis assay

Murine blood treated with heparin was centrifuged at 1,000g for 10 min and the

supernatant and the buffy coat were removed. Erythrocytes were washed three times with 0.9% saline solution and then resuspended to a concentration of 5% in saline containing 10% FBS (v/v). Red blood cells were then treated with serial dilutions (2 $\times$) of the peptides in a 96-well plate at 37 °C for 30 min. Control samples included a saline solution and 1% Triton X-100 as 0 and 100% haemolysis, respectively. Plates were centrifuged at 1,000g for 10 min and aliquots of the supernatant were diluted two times with saline solution. We measured the absorbance at 560 nm.

In vivo efficacy assay

MRSA bacteria (ATCC 33591) were grown at 37 °C with agitation for 12 h to a stationary phase. Cells were collected by centrifugation, washed twice with saline and resuspended to an A_{650} of 1.4. This suspension was diluted ten times in 5% mucin saline to a concentration of 3–5 $\times 10^7$ CFU ml⁻¹. For each peptide, five groups of eight mice (male BALB/c mice, 6 weeks old, about 20 g) per group were infected intraperitoneally with 0.5 ml of the above MRSA preparation (lethal dose). Forty-five minutes after infection, each group was treated intraperitoneally with a different dose of peptide, and a control group received vehicle only. We monitored mice for seven days; death was defined as the end point.

Received 8 February; accepted 24 May 2001.

- Koplan, J. P. *et al.* Preventing Emerging Infectious Diseases. A Strategy for the 21st Century (US Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, Georgia, 1998).
- Khurshid, M. A. *et al.* Staphylococcus aureus with reduced susceptibility to vancomycin. *Morbidity Mort. Week. Rep.* **48**, 1165–1167 (1999).
- Pinner, R. W. *et al.* Trends in infectious diseases mortality in the United States. *J. Am. Med. Assoc.* **275**, 189–193 (1996).
- Bong, D. T., Clark, T. D., Granja, J. R. & Ghadiri, M. R. Self-assembling organic nanotubes. *Angew. Chem. Int.* **40**, 988–1011 (2001).
- Ghadiri, M. R., Granja, J. R., Milligan, R. A., McRee, D. E. & Khazanovich, N. Self-assembling organic nanotubes based on a cyclic peptide architecture. *Nature* **366**, 324–327 (1993).
- Boman, H. G. Peptide antibiotics and their role in innate immunity. *Annu. Rev. Immunol.* **13**, 61–92 (1995).
- Oren, Z. & Shai, Y. Mode of action of linear amphipathic α -helical antimicrobial peptides. *Biopolymers* **47**, 451–463 (1998).
- Andreu, D. & Rivas, L. Animal antimicrobial peptides: an overview. *Biopolymers* **47**, 415–433 (1998).
- Oren, Z. & Shai, Y. Cyclization of a cytolytic amphipathic α -helical peptide and its diastereomer: effect on structure, interaction with model membranes, and biological function. *Biochemistry* **39**, 6103–6114 (2000).
- Ghadiri, M. R., Granja, J. R. & Buehler, L. Artificial transmembrane ion channels from self-assembling peptide nanotubes. *Nature* **369**, 301–304 (1994).
- Granja, J. R. & Ghadiri, M. R. Channel-mediated transport of glucose across lipid bilayers. *J. Am. Chem. Soc.* **116**, 10785–10786 (1994).
- Kim, H. S., Hartgerink, J. D. & Ghadiri, M. R. Oriented self-assembly of cyclic peptide nanotubes in lipid membranes. *J. Am. Chem. Soc.* **120**, 4417–4424 (1998).
- Kim, H. S. *Self-Assembling Cyclic D,L-Peptide Nanotubes as Synthetic Transmembrane Channels: From Chemical, Biological, and Practical Perspectives* Ch. 3–4. Thesis, The Scripps Research Institute, La Jolla, California (1999).
- Tosteson, M. T., Holmes, S. J., Razin, M. & Tosteson, D. C. Melittin lysis of red cells. *J. Membr. Biol.* **87**, 35–44 (1985).
- Sims, P. J., Wasgoner, A. S., Wang, C. H. & Hoffman, J. F. Studies on the mechanism by which cyanide dyes measure membrane potential in red blood cells and phosphatidylcholine vesicles. *Biochemistry* **13**, 3315–3330 (1974).
- Waggoner, A. S. Dye indicators of membrane potential. *Annu. Rev. Biophys. Bioeng.* **8**, 847–868 (1979).
- Loew, L. M. Characterization of potentiometric membrane dyes. *Adv. Chem. Ser.* **235**, 151–173 (1994).
- Friedrich, C. L., Moyses, D., Beveridge, T. J. & Hancock, R. E. W. Antibacterial action on structurally diverse cationic peptides on gram-positive bacteria. *Antimicrob. Agents Chemother.* **44**, 2086–2092 (2000).
- Amsterdam, D. in *Antibiotics in Laboratory Medicine* 3rd edn (ed. Lorian, V.) 53–105 (Baltimore, 1991).
- Seebach, D. *et al.* Cyclo- β -peptides: structure and tubular stacking of cyclic tetramers of 3-aminobutanoic acid as determined from powder diffraction data. *Helv. Chim. Acta* **80**, 173–182 (1997).
- Clark, T. D., Buehler, L. K. & Ghadiri, M. R. Cyclic β -peptide nanotubes as artificial transmembrane ion channels. *J. Am. Chem. Soc.* **120**, 651–656 (1998).
- National Committee for Clinical Laboratory Standards. *Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically* 4th edn M7-A4 (NCCLS, Villanova, Pennsylvania, 1997).

Acknowledgements

We thank M. Couto for his guidance and assistance with the *in vivo* studies. We are also grateful for the predoctoral fellowships from the Skaggs Research Institute and the Spanish Ministry of Education (S.F.-L.), and postdoctoral fellowships awarded by the Fulbright Program (G.V.L.), National Institutes of Health (D.A.W.), Swiss National Science Foundation (K.K.), and the Spanish Ministry of Education (M.D.). This work was supported by a grant to M.R.G. from the National Institutes of Health.

Correspondence and requests for materials should be addressed to M.R.G. (e-mail: ghadiri@scripps.edu).

RPA governs endonuclease switching during processing of Okazaki fragments in eukaryotes

Sung-Ho Bae, Kwang-Hee Bae, Jung-Ae Kim & Yeon-Soo Seo

National Creative Research Initiative Center for Cell Cycle Control, Samsung Biomedical Research Institute, Sungkyunkwan University School of Medicine, 300 Chunchun-Dong, Changan-Ku, Suwon, Kyunggi-Do 440-746, Korea

Extensive work on the maturation of lagging strands during the replication of simian virus 40 DNA suggests that the initiator RNA primers of Okazaki fragments are removed by the combined action of two nucleases, RNase HI and Fen1, before the Okazaki fragments join^{1–5}. Despite the well established *in vitro* roles of these two enzymes⁶, genetic analyses in yeast revealed that null mutants of RNase HI and/or Fen1 are not lethal^{7–9}, suggesting that an additional enzymatic activity may be required for the removal of RNA. One such enzyme is the *Saccharomyces cerevisiae* Dna2 helicase^{10–12}/endonuclease¹², which is essential for cell viability^{13,14} and is well suited to removing RNA primers of Okazaki fragments¹⁵. In addition, Dna2 interacts genetically and physically with several proteins involved in the elongation or maturation of Okazaki fragments^{10,16}. Here we show that the endonucleases Dna2 and Fen1 act sequentially to facilitate the complete removal of the primer RNA. The sequential action of these enzymes is governed by a single-stranded DNA-binding protein, replication protein-A (RPA). Our results demonstrate that the processing of Okazaki fragments in eukaryotes differs significantly from, and is more complicated than, that occurring in prokaryotes. We propose a novel biochemical mechanism for the maturation of eukaryotic Okazaki fragments.

The *S. cerevisiae* Dna2 mutant allele *Dna2Δ405N* encodes a protein lacking the amino-terminal 405 amino acids of Dna2 that is fully enzymatically active *in vitro*¹², but cells carrying this mutation show temperature-dependent growth (Fig. 1a). A screen of yeast genomic DNA libraries for one or more multicopy suppressors that corrected the growth defect yielded a clone containing a single intact open reading frame of *RFA2* that encodes the middle subunit of RPA (Fig. 1a), which has a relative molecular mass of 32,000 (*M_r* 32K). Subsequently, we found that genes encoding the large and small subunits of RPA (*RFA1* and *RFA3*) also suppressed the temperature-dependent lethality of *Dna2Δ405N* (Fig. 1a). This finding prompted us to examine the influence of RPA on the ability of Dna2 to remove the RNA-containing segment of Okazaki fragments in conjunction with displacement DNA synthesis (Fig. 1b). In the presence of eukaryotic polymerase δ (polδ) and its accessory proteins, replication factor-C (RFC) and proliferating cell nuclear antigen (PCNA), we observed that the addition of RPA (1 and 4 pmol) greatly stimulated the Dna2-catalysed cleavage of the 5'-labelled RNA segment (Fig. 1b, lanes 7 and 8). When we substituted *Escherichia coli* single-stranded-DNA-binding protein (SSB) for RPA in these reactions (which can also support displacement DNA synthesis by polδ (ref. 17)), the cleavage by Dna2 was stimulated at low levels (1 pmol) of SSB, but inhibited at higher levels (4 pmol; Fig. 1b, lanes 9 and 10), in contrast to the effects observed with RPA. When polδ was replaced with the Klenow fragment of *E. coli* DNA pol I, a polymerase with an ability to synthesize DNA that is unaffected by RPA or *E. coli* SSB, RPA markedly stimulated the cleavage reaction catalysed by Dna2, whereas SSB significantly inhibited Dna2 endonuclease activity (data not shown). Consistent with this, RPA stimulated the cleavage of the 5' single-stranded-DNA (ssDNA) overhang in a partial

duplex DNA substrate by Dna2 at physiological salt concentrations (Fig. 1c, lanes 3–6). In contrast, *E. coli* SSB inhibited this reaction (Fig. 1c, lanes 7–9). Dna2 cleaved the 3' ssDNA overhang poorly even in the presence of RPA (Fig. 1c, lanes 12–15). From these results, we conclude that RPA stimulates specifically the endonuclease activity of Dna2, and directs Dna2 to preferentially cleave the 5' ssDNA tail.

We next examined the effects of RPA on Dna2 endonuclease activity using three different flap-structured substrates, as illustrated in Fig. 1d. The addition of RPA (2.4 ng) markedly stimulated the Dna2-catalysed cleavage of the 27-nucleotide ssRNA–DNA (12 ribo- and 15 deoxyribonucleotides) chimaeric flap labelled at the 5' end (Fig. 1d, lanes 3 and 4). The level of products formed with 1 ng of Dna2 in the presence of 9.6 ng of RPA was comparable to that obtained with 20 ng of Dna2 (Fig. 1d, lanes 5 and 6). In the absence of the terminal 12-nucleotide ssRNA in the flap, RPA was weakly stimulatory (Fig. 1d, lanes 8–10) and more extensive cleavage was obtained only with excess amounts of Dna2 (Fig. 1d, lane 11). When we used the substrate with the 27-nucleotide ssDNA flap, the cleavage reaction was more efficient and RPA stimulation was greater than that observed with the 15-nucleotide flap DNA substrate (Fig. 1d, lanes 13 and 14). Therefore, although RPA does not interact with ssRNA (the binding affinity of RPA to ssRNA is 1,000-fold less than to ssDNA¹⁸), RPA-stimulated cleavage by Dna2 is sensitive to the overall length of the displaced flap regardless of DNA and RNA. Dna2 does not cleave RNA¹⁵, hence longer products were observed only with the chimaeric substrate.

To explore the possibility that RPA binding to ssDNA is necessary to recruit Dna2, we examined whether Dna2 forms a ternary complex with RPA-bound substrates using gel mobility assays. Dna2 alone formed a complex efficiently with the 69-nucleotide 5' ssDNA overhang substrate (Fig. 2a, lanes 2–4). In the presence of RPA, a new complex was detected that migrated slower than the RPA–substrate or Dna2–substrate complexes (Fig. 2a, lanes 6–8). The amount of the ternary complex formed in the presence of RPA was greater (~3-fold) than the amount of complex formed with Dna2 alone (Fig. 2a, compare lanes 2–4 with 6–8), indicating that the complex formation is cooperative. In contrast, the addition of SSB prevented Dna2 from forming a complex with the substrate DNA (Fig. 2a, lanes 9–12). When we used a flap-structured DNA containing a 35-nucleotide ssDNA tail, RPA and Dna2 also formed a ternary complex (Fig. 2b, lanes 1–6) as efficiently as that observed with the 69-nucleotide 5' ssDNA overhang. We found that any ssDNA tail longer than 27 nucleotides supported the ternary complex formation and RPA stimulation of Dna2 cleavage with similar efficiencies (data not shown). This result also indicates that the ssDNA-binding activity of RPA is critical for stimulation of Dna2. In support of this, the addition of 10-fold molar excess of the 48-nucleotide ssDNA flanked by partial duplexes at both ends, which is hardly cleaved by Dna2 (ref. 15), abolished the RPA-stimulation of the Dna2-catalysed cleavage (data not shown).

The ternary complex seems to be an intermediate structure that is eventually converted to products, as the addition of 2 mM Mg²⁺ caused the complex to dissociate (Fig. 2b, lanes 7 and 8). This was dependent on functional Dna2: ternary complex formation with Dna2D657A, a mutant enzyme devoid of endonuclease activity¹³, did not disassemble in the presence of Mg²⁺ (data not shown). With low levels of Dna2 (0.1 or 0.2 ng), the RPA–DNA complex that was formed disappeared at a rate proportional to the amount of Dna2 added, with the concomitant increased formation of cleavage products that migrated faster than the substrate DNA (Fig. 2b, lanes 9–11). The flap DNA remaining after Dna2-catalysed cleavage of RPA-bound ssDNA tails varied in length from 5 to 7 nucleotides (see Fig. 3b, lanes 10–12), a size insufficient to support the stable association of RPA with ssDNA¹⁹. This suggests that RPA should rapidly recycle in the presence of Dna2 and catalytically stimulate

the endonuclease activity of Dna2. To confirm this idea, we examined the effect of increasing levels (5–200 fmol) of RPA on the cleavage reaction with a fixed amount of Dna2 (20 fmol) and substrate (100 fmol) (Fig. 2c). In the absence of RPA, the extent of DNA cleavage catalysed by Dna2 alone did not exceed 20 fmol even after the longest incubation time (30 min). The addition of limited amounts (5–10 fmol) of RPA, far less than that of the substrate, markedly stimulated the reaction, and the cleavage rate increased in proportion to the amount of RPA added (Fig. 2c). The addition of

RPA above the level (100 fmol) of the substrate failed to further increase the cleavage rate of Dna2 (Fig. 2c). These results confirm that RPA accelerates the rate of substrate cleavage by Dna2 in a catalytic fashion. Consistent with a synergistic complex formation between Dna2 and RPA that suggests a protein–protein interaction, Dna2 interacted with RPA in yeast two-hybrid assays (data not shown) and in enzyme-linked immunosorbent assays (ELISAs). However, no interaction was detected between Dna2 and *E. coli* SSB (Fig. 2d).

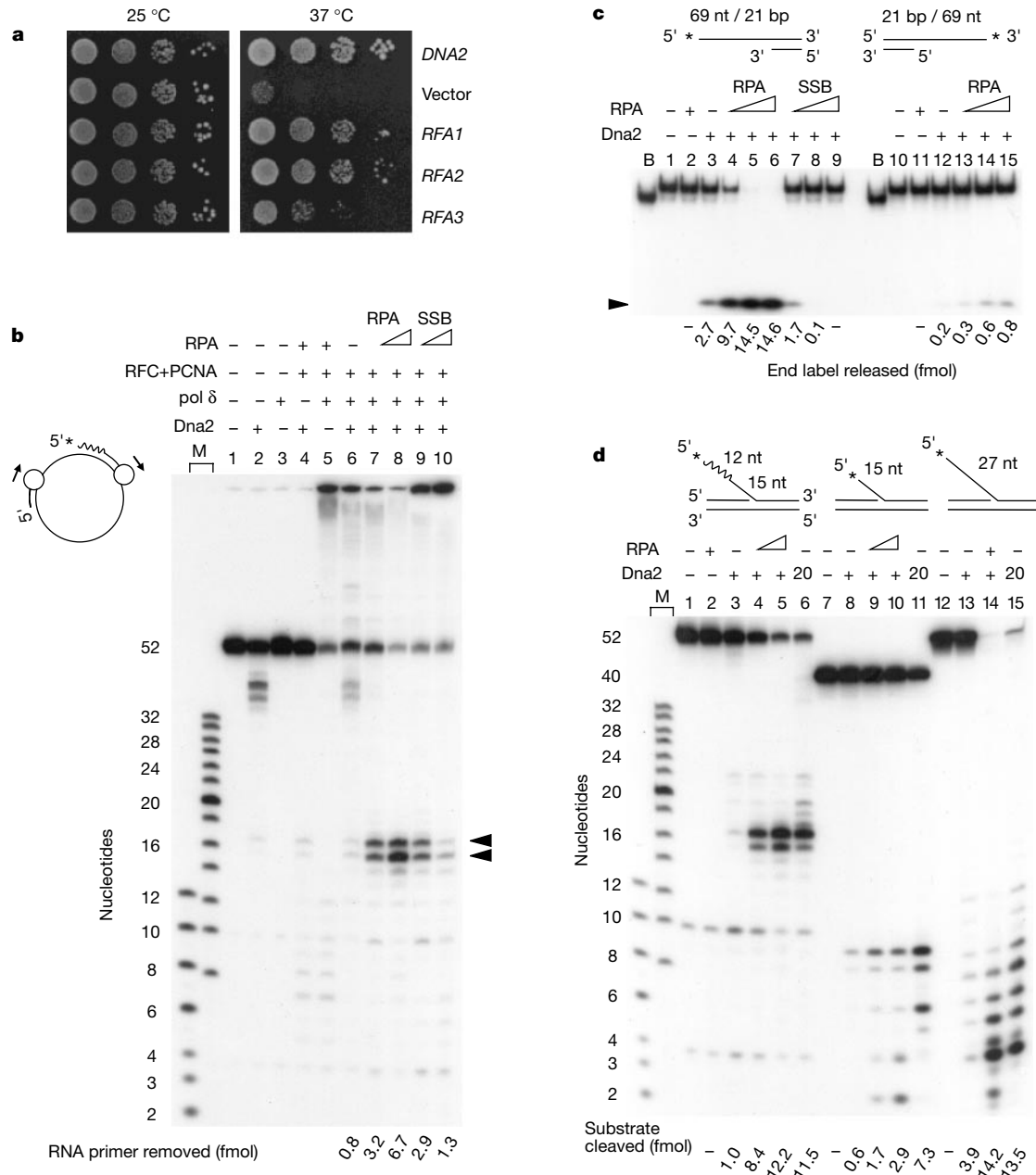


Figure 1 RPA has a genetic interaction with Dna2 and stimulates endonuclease activity of Dna2. **a**, A yeast strain carrying the *Dna2Δ405N* mutation (a derivative of YPH499 (*MATa, ade2, ura3, lys2, trp1, his3, leu2*)), was transformed with vector alone (pRS325; New England Biolab) or vector containing *DNA2*, *RFA1*, *RFA2* or *RFA3*. Aliquots containing transformed cells (5×10^4 , 10^3 , 10^2 or 10^1) were spotted onto SD-Leu plates and grown at 25 or 37 °C in duplicate. Expression of each gene is driven by its own promoter. **b**, RPA-specific stimulation of Dna2-catalysed cleavage of the RNA segment during displacement DNA synthesis. The reaction condition and product analysis were as described previously¹⁵. Arrowheads indicate major cleavage products formed after

incubation with Dna2. M, marker; asterisk, the position of the ³²P label in the substrate. Circles indicate the pol δ holoenzyme. **c**, Standard reaction mixtures containing 125 mM NaCl were incubated with increasing amounts of RPA or *E. coli* SSB (5, 20 or 80 fmol) and 0.5 ng of Dna2 at 37 °C for 5 min. The reaction products were analysed with a low-resolution gel¹⁵. **b**, boiled substrate control; nt, nucleotides; bp, base pairs. Arrowhead indicates cleavage products. **d**, The substrate on the left contains a 12-nucleotide terminal RNA segment (wavy line). Standard reactions containing 125 mM NaCl were incubated with Dna2 (1 ng) and RPA (2.4 ng = 20 fmol). Reactions shown in lanes 6, 11 and 15 contained 20 ng of Dna2. Increasing amounts of RPA were 20 and 80 fmol.

On the basis of the experimental evidence described above, we thought that RPA could mediate the sequential action of Dna2 and Fen1 in the removal of primer RNA present in Okazaki fragments. Therefore we investigated the effect of RPA on junction cleavage by Fen1. Junction-specific cleavage by Fen1 was inhibited by increasing amounts of RPA (Fig. 3a, lanes 3–6) but not by Dna2 (Fig. 3a, lanes 7 and 8). The cleavage activity of Fen1 was nearly completely inhibited in the presence of 4 ng RPA (33 fmol), an amount sufficient to bind to all of the added substrate (Fig. 3a, lane 6). This result suggests that dissociation of RPA from the 5' ssDNA tail is an essential step before Fen1 can remove the remaining flap structure.

To determine whether Dna2 could relieve the inhibitory effect of RPA on Fen1, the substrate labelled at the 3' end of the strand containing the 5' ssDNA flap was incubated with Dna2, Fen1 and RPA, and the length of ssDNA flap remaining after cleavage was analysed with a high-resolution sequencing gel. In the absence of RPA, Fen1 and Dna2 seemed to act independently of each other (Fig. 3b, lanes 2–6). In the presence of RPA (4 ng), however, cleavage of the flap DNA by Fen1 was markedly (>10-fold) inhibited (Fig. 3b, compare lanes 2 and 3 with 8 and 9), whereas its cleavage by Dna2 was significantly stimulated (Fig. 3b, compare lanes 4 and 10). When all three proteins were present, the shortened flap product generated by Dna2 was processed further by Fen1 to a nicked duplex product identical in length to that generated by the action of Fen1 alone (Fig. 3b, lanes 11 and 12).

To establish the order of the cleavage reaction, we performed

time-course analyses with three different levels of Dna2 (0, 0.05 and 0.2 ng) and fixed levels of Fen1 (5 ng) and RPA (4 ng; Fig. 3c). Dna2 and Fen1 each generated distinguishable major products, as shown in Fig. 3b. The intensity of these bands was measured at each time point and plotted in Fig. 3c. The products generated by Dna2 appeared more rapidly than those by Fen1; moreover, the product formed by Dna2 decreased with time, whereas the products formed by Fen1 increased in proportion to the incubation time (Fig. 3c). Most importantly, levels of products formed by Fen1 action increased in response to the amount of Dna2 even though these reaction mixtures contained a fixed concentration of Fen1 (Fig. 3c, open symbols). These results indicate that the action of Dna2 generates the substrate used by Fen1 and governs the initial rate of Fen1 activity, indicating that, in the presence of RPA, Dna2 acts before Fen1. The dependence of Fen1 activity on Dna2 supports the idea that Dna2 and Fen1 act in a sequential and coordinated manner in the same pathway.

Finally, we investigated whether the products generated by the action of Dna2 and Fen1 in the presence of RPA are efficiently ligated by yeast DNA ligase I. As shown in Fig. 3d, the limited amounts of cleavage products formed by Dna2 were ligated to form a full-length (51-nucleotide) oligonucleotide equivalent to the template size, although the efficiency of ligation was very poor and unaffected by RPA (Fig. 3d, lanes 3–5, see also the bottom of the figure for an overexposed image). In contrast, the products formed by Fen1 were efficiently ligated, resulting in the formation of the 51-nucleotide oligonucleotide (Fig. 3d, lane 7). The addition of RPA

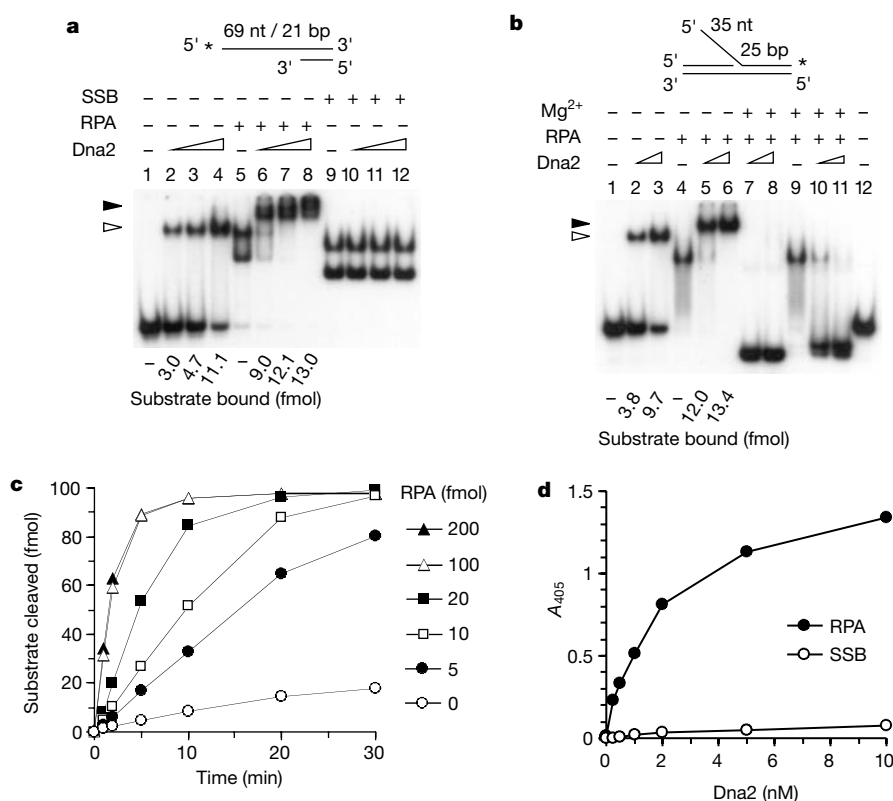


Figure 2 RPA and Dna2 form a ternary complex with 5'-tailed DNA substrates. **a**, Increasing amounts of Dna2 (5, 10 and 20 ng) were incubated with the ssDNA in the presence (+) or absence (-) of 20 fmol of RPA or *E. coli* SSB. Arrowheads indicate the positions of Dna2–DNA (open) and Dna2–RPA–DNA (filled) complexes. nt, nucleotides; bp, base pairs; asterisk, the position of the ³²P label in the substrate. **b**, RPA (2.4 ng = 20 fmol) and increasing amounts (10 and 20 ng) of Dna2 were used for complex formation with or without 2 mM MgCl₂. In lanes 10 and 11, limited amounts of Dna2 (0.1 and 0.2 ng) were used to monitor the fate of the RPA–DNA complex. Arrowheads indicate the

migration positions of Dna2–DNA (open) and Dna2–RPA–DNA (filled) complexes. **c**, The substrate used was the same as in **b**, but labelled at the 5' end of the flap. Reaction mixtures (160 μl) containing 125 mM NaCl, 100 fmol of the DNA substrate, 3.4 ng (20 fmol) of Dna2 and increasing amounts of RPA were preincubated at 37 °C for 5 min. Nuclease activation was initiated by the addition of 2 mM Mg²⁺, and 20-μl aliquots were withdrawn at each time point. **d**, Detection of Dna2–RPA complex formation by ELISA. The mean of three experiments is shown. A₄₀₅, absorbance at wavelength 405 nm.

inhibited formation of ligated DNA, in keeping with the observation that RPA blocks the Fen1 junction cleavage reaction (Fig. 3d, lane 8). However, the addition of Dna2 to the mixture containing both Fen1 and RPA fully restored the ligation reaction (Fig. 3d, lane 9).

On the basis of our results, we propose a novel mechanism of processing of Okazaki fragments in eukaryotes. This model comprises several steps (Fig. 4). (1) The RNA containing the 5' terminus of an Okazaki fragment is rendered single-stranded by displacement DNA synthesis catalysed by pol δ (ref. 15). (2) RPA rapidly forms an

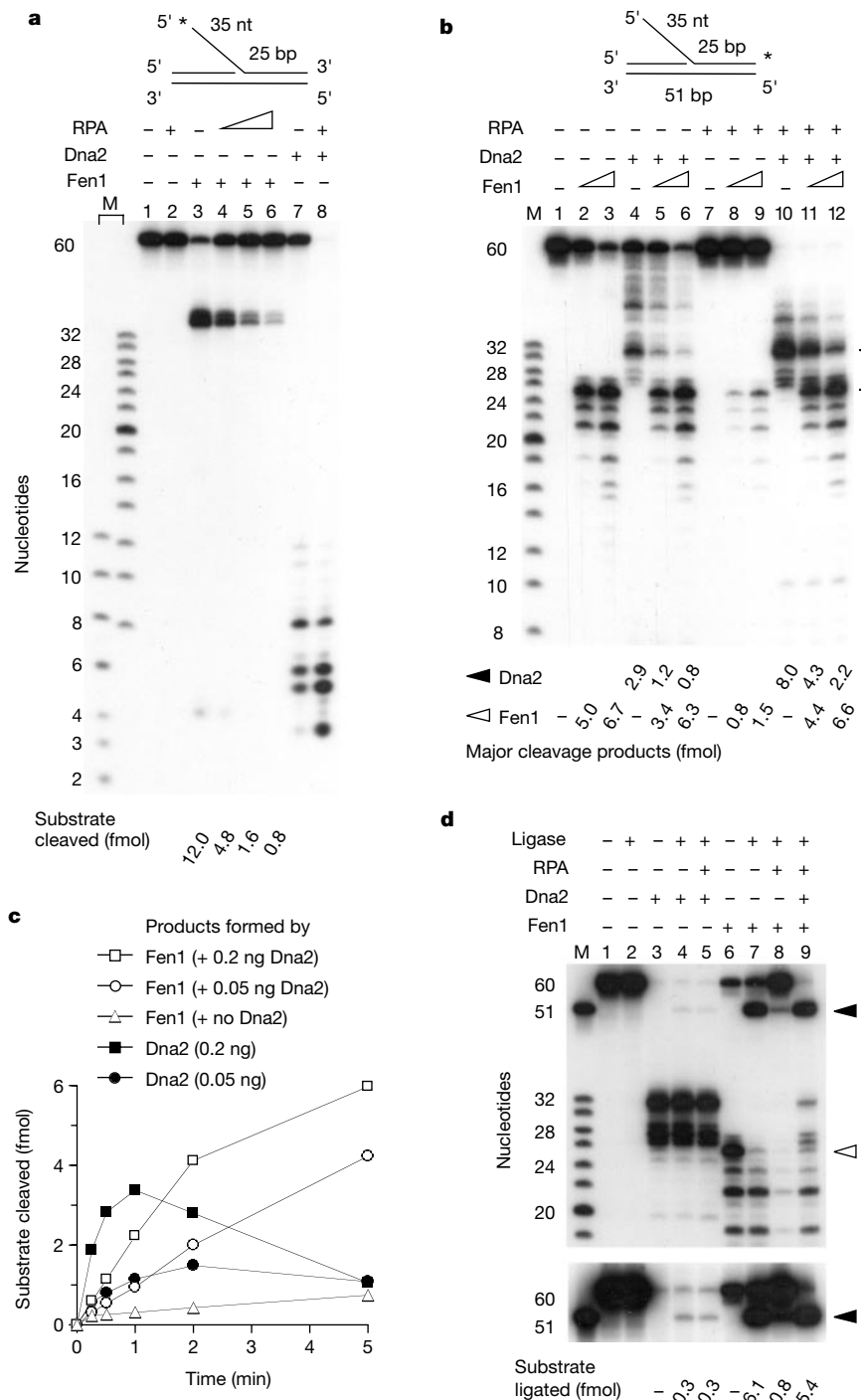


Figure 3 RPA mediates the sequential and ordered action of Dna2 and Fen1 in the processing of Okazaki fragments. The flap substrates were labelled at the 5' (**a**) or 3' (**b**, **c** and **d**) end of the downstream primers. The first 10 nucleotides of the downstream primer in **d** were substituted for ribonucleotides to simulate *in vivo* substrates. Standard reactions containing 50 mM NaCl were carried out with or without RPA (4 ng), Dna2 (0.2 ng), Fen1 (5 ng) and/or DNA ligase I (1 μl). Reactions were preincubated at 37 °C for 5 min and initiated by the addition of 2 mM Mg²⁺; products were analysed in high-resolution sequencing gels¹⁵. **a**, Increasing levels (1, 2 and 4 ng) of RPA were added as shown. nt, nucleotides; bp, base pairs; M, marker; asterisk, ³²P label. **b**, Reactions were

performed with increasing amounts of Fen1 (2 and 5 ng). Arrowheads indicate the major cleavage products formed by Dna2 (filled) and Fen1 (open). **c**, Amounts of the major products generated by Dna2 and Fen1 (indicated by arrowheads in **b**) from reactions containing constant levels of RPA (4 ng) and Fen1 (5 ng), but different amounts of Dna2. **d**, Reactions were carried out at 37 °C for 5 min in the presence of 1 mM ATP. The open arrowhead denotes the major product formed from the junction cleavage by Fen1; the filled arrowhead indicates the expected size (51 nucleotides) of the ligated product. An overexposed autoradiography of the top part of the gel is shown at the bottom.

initial complex with the nascent flap structure and (3) then recruits Dna2 to form a ternary complex. This leads to the initial cleavage of RNA-containing segments by Dna2. This proposal could account for the fact that Dna2 has been shown to be essential^{16,20}, whereas Fen1 is not. The binding of RPA to the displaced end of RNA-containing Okazaki segments does not allow Fen1 to access the substrate. In the absence of Dna2, the processing of Okazaki fragments is blocked and cells are not viable. (4) The cleavage of the displaced RNA-containing flap by Dna2 results in the disassembly of the ternary complex, allowing both Dna2 and RPA to recycle. The remaining short flap DNA can be processed further either by Fen1, which is loaded onto the DNA through protein–protein interactions with PCNA²¹ (Fen1 dependent) or alternatively by other nucleases, possibly Exo1 or Dna2 itself (Fen1 independent). The fact that ligation products were formed, although poorly, with Dna2 alone (Fig. 3d) also raises the possibility that Dna2 by itself could constitute the simplest Fen1-independent processing pathway. (5) Finally, the resulting nick is sealed by DNA ligase I.

This model does not necessarily rule out the possibility that two nucleases, Dna2 and Fen1 with RNase HI constitute parallel

pathways under some circumstances. If displacement DNA synthesis by pol δ were inefficient, for example, Okazaki fragments could be susceptible to the action of RNase HI and Fen1 before the upstream polymerase arrives. In this case, the 5'-to-3' exonuclease activity of Fen1 could contribute to the removal of the RNA-initiated primer with the aid of RNase HI (ref. 6).

Our findings that RPA plays a role in Okazaki fragment processing *in vivo* are supported by two independent observations. First, a mutation in *DNA2* was identified during a synthetic lethal screen with *rfa1Y29H*, a temperature-sensitive mutant allele of *RPA1* (S. Brill, personal communication), indicating a functional interaction between Dna2 and RPA. Second, the 32K subunit of RPA was crosslinked to RNA–DNA primers in the lagging strand of replicating simian virus 40 chromosomes²². This crosslinking was observed only with early RNA–DNA primer intermediates and was not detected with mature lagging strand products, consistent with our *in vitro* finding that RPA first binds to the 5' end of RNA–DNA primers displaced by pol δ and subsequently dissociates from the flap.

The proposed model could account for several complicated genetic observations made on Okazaki fragment maturation. First, the inactivation of either Fen1, Exo1 or Mre11 did not lead to cell death. However, the inactivation of any two of these enzymes resulted in cell death^{14,23}, probably owing to the inefficient generation of nicked duplex DNA. The overexpression of Exo1 in the absence of Fen1 restored the growth of mutant cells at the non-permissive temperature²³. Second, an *in vivo* study demonstrated that repair induced by double-strand breaks, which requires replication of both leading and lagging DNA strands, was 50% lower in Fen1-deficient strains than in wild-type strains²⁴, suggesting that the remaining 50% of repairs are carried out through a Fen1-independent pathway. Third, genetic analyses of *rad27* mutant alleles also suggest the existence of a Fen1-independent pathway²⁵. The interaction of nuclease-deficient Rad27-n with PCNA freezes the mutant protein at its normal functioning position, preventing an alternative nuclease from binding to the incompletely processed DNA. This can be reversed by introducing a PCNA-binding defective mutation in Rad27-n (Rad27-n,p).

Why has this complicated mechanism for the removal of primer RNA evolved in eukaryotes? Perhaps the mechanism permits eukaryotic cells to remove the DNA beyond the RNA–DNA junction. The removal of the entire RNA–DNA primer region synthesized by the pol α –primase complex, which lacks a proof-reading activity, may abrogate the need to correct any errors inserted by pol α . Our *in vitro* data are consistent with this possibility as flap structures longer than 27 nucleotides, comparable to the size (~35 nucleotides) of the entire RNA–DNA primer, are removed most efficiently by the combined action of Dna2, RPA and Fen1. Further experiments are underway to test how accurately the details of our *in vitro* model reflect the processing of Okazaki fragments *in vivo*. □

Methods

Oligonucleotides and substrate preparations

All oligonucleotides used for the construction of DNA and RNA substrates were synthesized commercially and gel-purified before use¹⁵. The procedures for substrate preparation and labelling at either end of the substrate are as described previously^{12,15}. Substrate structures and positions of radioisotopic labels are described in the figures. The primed partial duplex substrate (Fig. 1b) and the 5'- or 3'-overhang partial duplex substrates (Figs 1c and 2a) are described elsewhere¹⁵. In order to construct the flap-structured substrates with different 5'-tail lengths, both downstream (either 40, 52 or 60 nucleotides) and upstream primers were annealed to the template primer. The sequences of these primers used are as follows: the 60-nucleotide downstream primer, 5'-GGTAAACGCGAACAATTCAGCGGCTTTAACCAGGACGCTCGACGCCATTAATAAT GTTTTC-3'; the upstream primer, 5'-CCGTTAGCAGTTTCGCTTGTGCCTA-3' (25 nucleotides); and the template primer, 5'-GGAAAACATTATTAATGGCGTCGAGCTAG GCACAAGGCGAAGTCTAAGCG-3' (51 nucleotides). The two shorter downstream primers (52 and 40 nucleotides) lacked the first 8 and 20 nucleotides, respectively, of the 60-nucleotide downstream primer.

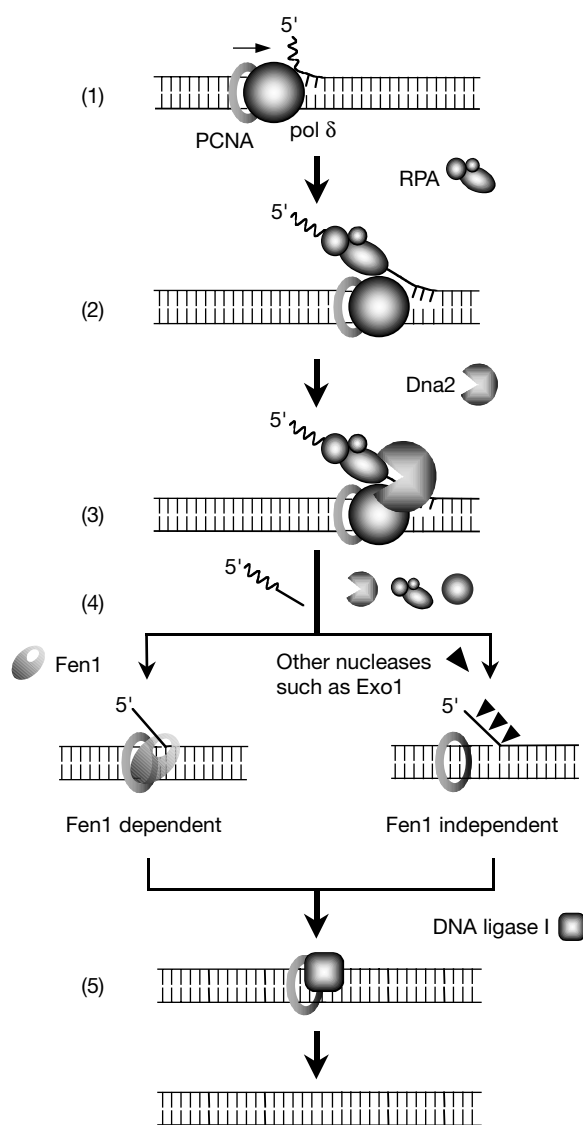


Figure 4 A model of processing of Okazaki fragments in eukaryotes. The wavy line followed by a straight line denotes the primer RNA and DNA, respectively. The direction of Okazaki fragment extension by pol δ is indicated by an arrow in (1). RFC is not shown for simplicity. See text for details.

Proteins and enzymes

Escherichia coli SSB was purchased from Amersham Pharmacia Biotech. The recombinant Dna2 protein, PCNA, RFC and Fen1 were prepared as described previously^{12,15,26}. RPA was purified from a protease-deficient yeast strain BJ2168 (*MATa*, *ura3-52*, *trp1-Δ63*, *leu2-Δ*, *prb1-1122*, *pep4-3*, *prc1-407*, *gal2*) as described²⁷. DNA ligase I was partially purified from BJ2168 as described²⁸. *Schizosaccharomyces pombe* DNA pol δ (ref. 29) was a gift from J. Hurwitz.

Endonuclease assays

Standard reaction mixtures (40 μl) contained 50 mM Tris-HCl buffer (pH 7.8), 2 mM MgCl₂, 2 mM dithiothreitol, 0.25 mg ml⁻¹ bovine serum albumin (BSA) and 15 fmol of substrate. Other variations are indicated in figure legends. Reactions were carried out at 37 °C for 5 min unless otherwise indicated and the products were analysed either in high-resolution denaturing gels¹⁵ for visualization of the products or in low-resolution gels¹² for quantification purposes only. The markers were as described¹⁵, and their sizes are indicated in each figure.

Gel mobility shift assay

Reaction mixtures (40 μl) contained 50 mM Tris-HCl buffer (pH 7.8), 2 mM dithiothreitol, 0.25 mg ml⁻¹ BSA, 125 mM NaCl and 15 fmol of substrate. Reactions were incubated on ice for 10 min, followed by further incubation at 37 °C for 5 min. Glycerol and bromophenol blue were added to 10% and 0.05%, respectively, and products were separated for 1 h at 100 V on a 5% acrylamide gel in 0.5× TBE. The amount of complexes formed was measured with a PhosphorImager.

ELISA assay

The ELISA assay was performed as described³⁰. Microtitre wells coated with RPA or *E. coli* SSB were incubated with 100 μl of increasing amounts of Dna2 (0–1.7 ng μl⁻¹) in 50 mM Tris buffer (pH 7.5), 125 mM NaCl, 100 μg ml⁻¹ BSA and 0.02% NP-40 for 30 min at 25 °C. Dna2–RPA complexes were detected using rabbit polyclonal antibody against Dna2 (ref. 12) as the primary antibody and the ABTS substrate (Roche) for the colouring reaction. Absorbance readings were corrected by subtracting the background signal in BSA-coated wells.

Received 1 November 2000; accepted 4 June 2001.

1. Challberg, M. D. & Kelly, T. J. Animal virus DNA replication. *Annu. Rev. Biochem.* **58**, 671–717 (1989).
2. Hurwitz, J., Dean, F. B., Kwong, A. D. & Lee, S. H. The *in vitro* replication of DNA containing the SV40 origin. *J. Biol. Chem.* **265**, 18043–18046 (1990).
3. Waga, S. & Stillman, B. The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* **67**, 721–751 (1998).
4. Lieber, M. R. The FEN1 family of structure-specific nucleases in eukaryotic DNA replication, recombination and repair. *Bioessays* **19**, 233–240 (1997).
5. Bambara, R. A., Murante, R. S. & Hendericksen, L. A. Enzymes and reactions at the eukaryotic DNA replication forks. *J. Biol. Chem.* **272**, 4647–4650 (1997).
6. Waga, S. & Stillman, B. Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication *in vitro*. *Nature* **369**, 207–212 (1994).
7. Frank, P., Braunshofer-Reiter, C. & Wintersberger, U. Yeast RNase H(35) is the counterpart of the mammalian RNase HI, and is evolutionarily related to prokaryotic RNase HII. *FEBS Lett.* **421**, 23–26 (1998).
8. Reagan, M. S., Pittinger, C., Siede, W. & Friedberg, E. C. Characterization of a mutant strain of *Saccharomyces cerevisiae* with a deletion of the RAD 27 gene, a structural homolog of the RAD2 nucleotide excision repair gene. *J. Bacteriol.* **177**, 364–371 (1995).
9. Sommers, C. H. *et al.* Conditional lethality of null mutations in RTH1 that encodes the yeast

counterpart of a mammalian 5' to 3'-exonuclease required for lagging strand DNA synthesis in reconstituted systems. *J. Biol. Chem.* **270**, 4193–4196 (1995).

10. Budd, M. E. & Campbell, J. A yeast replicative helicase, Dna2 helicase, interacts with yeast FEN1 nucleases in carrying out its essential function. *Mol. Cell. Biol.* **17**, 2136–2142 (1997).
11. Budd, M. E. & Campbell, J. A yeast gene required for DNA replication encodes a protein with homology to DNA helicases. *Proc. Natl Acad. Sci. USA* **92**, 7642–7646 (1995).
12. Bae, S. H. *et al.* Dna2 of *Saccharomyces cerevisiae* possesses a single-stranded DNA-specific endonuclease activity that is able to act on double-stranded DNA in the presence of ATP. *J. Biol. Chem.* **273**, 26880–26890 (1998).
13. Lee, K. W. *et al.* The endonuclease activity of the yeast Dna2 enzyme is essential *in vivo*. *Nucleic Acids Res.* **28**, 2873–2881 (2000).
14. Budd, M. E., Choe, W.-C. & Campbell, J. The nuclease activity of the yeast Dna2 protein, which is related to the RecB-like nucleases, is essential *in vivo*. *J. Biol. Chem.* **275**, 16518–16529 (2000).
15. Bae, S. H. & Seo, Y. S. Characterization of the enzymatic properties of the yeast Dna2 helicase/endonuclease suggests a new model for Okazaki fragment processing. *J. Biol. Chem.* **275**, 38022–38031 (2000).
16. Kang, H. Y. *et al.* Genetic analyses of *Schizosaccharomyces pombe* *dna2*⁺ reveal that Dna2 plays an essential role in Okazaki fragment metabolism. *Genetics* **155**, 1055–1067 (2000).
17. Kenny, M., Lee, S.-H. & Hurwitz, J. Multiple functions of human single-stranded-DNA binding protein in simian virus 40 DNA replication: single-strand stabilization and stimulation of DNA polymerases α and δ. *Proc. Natl Acad. Sci. USA* **86**, 9757–9761 (1989).
18. Wold, M. S. Replication protein A: a heterotrimeric, single-stranded DNA binding protein required for eukaryotic DNA replication. *Annu. Rev. Biochem.* **66**, 61–92 (1997).
19. Ifode, C. & Borowicz, J. A. Denaturation of the simian virus 40 origin of replication mediated by human replication protein A. *Mol. Cell. Biol.* **17**, 3876–3883 (1997).
20. Budd, M. E., Choe, W.-C. & Campbell, J. Dna2 encodes a DNA helicase essential for replication of eukaryotic chromosomes. *J. Biol. Chem.* **270**, 26766–26769 (1995).
21. Gomes, X. V. & Burgers, P. M. J. Two modes of FEN1 binding to PCNA regulated by DNA. *EMBO J.* **19**, 3811–3821 (2000).
22. Mass, G., Nethanel, T. & Kaufmann, G. The middle subunit of replication protein A contacts growing RNA–DNA primers in replicating simian virus 40 chromosomes. *Mol. Cell. Biol.* **18**, 6399–6407 (1998).
23. Tishkoff, D. X. *et al.* Identification and characterization of *Saccharomyces cerevisiae* EXO1, a gene encoding an exonuclease that interacts with MSH2. *Proc. Natl Acad. Sci. USA* **94**, 7487–7492 (1997).
24. Holmes, A. M. & Haber, J. Double-strand break repair in yeast requires both leading and lagging strand DNA polymerases. *Cell* **96**, 415–424 (1999).
25. Gary, R. *et al.* A novel role in DNA metabolism for the binding of Fen1/Rad27 to PCNA and implications for genetic risk. *Mol. Cell. Biol.* **19**, 5373–5382 (1999).
26. Bae, S. H. & Seo, Y. S. *In vitro* evidence that purified yeast Rad27 and Dna2 are not stably associated with each other suggests that an additional protein(s) is required for a complex formation. *J. Biochem. Mol. Biol.* **33**, 155–161 (2000).
27. Brill, S. J. & Stillman, B. Yeast replication factor-A functions in the unwinding of the SV40 origin of DNA replication. *Nature* **342**, 92–95 (1989).
28. Tomkinson, A. E., Tappe, N. J. & Friedberg, E. C. DNA ligase I from *Saccharomyces cerevisiae*: physical and biochemical characterization of the CDC9 gene product. *Biochemistry* **31**, 11762–11771 (1992).
29. Zuo, S. *et al.* Structure and activity associated with multiple forms of *Schizosaccharomyces pombe* DNA polymerase δ. *J. Biol. Chem.* **275**, 5153–5162 (2000).
30. Brosh, R. M. *et al.* Replication protein A physically interacts with the Bloom's syndrome protein and stimulates its helicase activity. *J. Biol. Chem.* **275**, 23500–23508 (2000).

Acknowledgements

We thank J. Hurwitz, S. MacNeill and A.-K. Bielinsky for critical reading of the manuscript. This work was supported by a grant from the Creative Research Initiatives of the Korean Ministry of Science and Technology given to Y.-S.S.

Correspondence and requests for materials should be addressed to Y.-S.S. (e-mail: ysseo@med.skku.ac.kr).

Contacts

Publisher: Fabien Savenay

Editor: Paul Smaglik

Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:

Leonie Welts (4954)

European Manager:

Nevin Bayoumi (4978)

European Sales Consultant:

Jonathan Bull (4973)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),

Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Sille Opstrup (4994)

Production Manager:

Billie Franklin

To send films and materials use

London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau

Tel +33 (0) 1 43 20 16 51

Fax +33 (0) 1 43 20 51 52

e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Amanda Wren

Tel +34 91 391 3835

Fax +34 91 319 0091

e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel +49 89 54 90 57 11/-2

Fax +49 89 54 90 57 20

e-mails: p.phelan@nature.com

k.turner@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: classified@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),

3-6 Ichigaya Tamachi Shinjuku-Ku,

Tokyo 162

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.cowan@nature.com

Japan Manager:

Kate Cowan

More data, more potential

After the public and private human-genome teams unveiled their respective works-in-progress earlier this year, there was some concern that interest in genetics and genomics would recede. Fortunately, the opposite has happened. Two recent pieces of news show that commitment to — and opportunities in — genomics and genetics shows no sign of letting up.

This month many of the same players involved in the public effort — the Wellcome Trust, the US National Institutes of Health and the Whitehead Institute — announced plans to map clusters of genetic variations that are associated with each other (see *Nature* 412, 105; 2001). Like the effort to map single-nucleotide polymorphisms, the \$60-million effort will also be funded by biotech companies and pharmaceutical companies.

Also, the European Commission's investment in the European Molecular Biology Laboratory (EMBL) earlier this year (see *Nature* 411, 229; 2001) looks as though it will soon yield dividends. This month, Janet Thornton, a structural-biology professor at University College London, was named research director of the European Bioinformatics Institute near Cambridge in England. And Nadia Rosenthal, an associate molecular-biology professor at Harvard University, was named as head of the mouse biology programme at the European Mouse Mutant Archive near Rome. The news is doubly good — first because it shows that women scientists are increasingly rising to leadership positions, and second, because both women have been charged with expanding each institution from service provider to research organization.

With the right mixture of luck and diligence, the launch of the mapping project and the expansion of the EMBL will yield more scientific insight — and more employment opportunities.

Paul Smaglik

Naturejobs editor



Contents

REGIONS

Hong Kong gears up for a biotech revolution

p4

WWW.NATURE.COM/ NATUREJOBS

Career centre

Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



ROBERT HARDING

Putting the pieces together Hong Kong

For the past 15 years, Hong Kong has been luring top international talent in the biological sciences to its expanding university system with high salaries and quality equipment (see “Big salaries, little grants”). It has an immense stock of bright researchers from mainland China who are more than willing to work in Hong Kong (see “Banking on China”). It is internationally connected by airports and English-speaking ability. It has a lot of money, mostly from commerce and banking, which could provide venture capital.

Yet it is struggling in its efforts to forge a biotechnology industry. “All the elements to make biotech are here, but nothing has happened,” says Frederick Leung of Hong Kong University (HKU).

Leung says that one major stumbling block is the lack of private investment available for long-term projects. Investors want to make money in two years, says Nancy Ip, director of the Biotechnology Research Institute (BRI) at Hong Kong University of Science and Technology (HKUST).

“Investors’ attitudes have started to change,” says David Miller-Martini, the BRI’s business-development manager. “But they still need to be convinced.”

Most support for biotechnology — HK\$800 million (US\$100 million) between 1994 and 1999 — comes from the government. Most of that goes through the Innovation and Technology Commission (ITC) into Hong

Kong’s three main universities — HKU, the Chinese University of Hong Kong and HKUST — and some to its recently upgraded polytechnics such as the City University of Hong Kong.

A wide range of promising research is pursued at these. But in the absence of a biotech market at home, “there’s no industry to pick up technology and run with it,” says Miller-Martini. If the technology is to find application, it will probably be overseas.

That is not to say that universities do not do applied research. Over half of the funding to HKU’s zoology department is from industry, says department chair Danny Chan. Chan has helped to transform the department by introducing research programmes in biotechnology, food technology, environmental science and agriculture. There are also larger-scale, longer-term projects, but mostly with government support. Like many others, Michael Yang from the City University of Hong Kong is hoping that by focusing on biomedical problems of regional importance he can find a niche in the crowded world of microarrays. The Applied Research Centre for Genomic Technology, which Yang directs, has impressive facilities for making biochips for diagnosing diseases and detecting differences in people’s genetic make-up. Yang says his microarray technology will also elucidate the mechanisms of traditional Chinese medicines and the avian ‘flu virus.

Outside of the university, some companies are also trying to find their way. Genetel Pharmaceuticals is developing DNA microarrays for detecting the genes behind obesity and osteoporosis. Hong Kong DNA Chips is also trying to break into the ‘lab-on-a-chip’ business to simplify detection and diagnostics. To pay the high start-up costs for its research and

Danny Chan has helped to transform Hong Kong University’s zoology department by introducing fresh research programmes.



Banking on China

In the 1990s, many were worried about what would happen when the British turned Hong Kong over to China.

But since becoming a 'special administrative region' of China in 1997, Hong Kong has retained a lot of its independence. Attitudes have changed from fear to need: "The mainland is the key for Hong Kong to survive in biotech," says Danny Ngia, whose company, Chinetek Scientific, has been selling scientific imaging services throughout China. This sentiment is widespread in Hong Kong.

China provides about half of Hong Kong's researchers. "We expect to shift increasingly to

mainland researchers," says Jeffrey Wong of the Hong Kong University of Science and Technology (HKUST). Chinese postdocs will often work for half the price of others, says another HKUST professor.

More than anything, China is a market that could support industry. It accounted for 68% of Hong Kong's pharmaceutical and medical exports last year.

But will China some day swallow Hong Kong whole? "I'm optimistic for Hong Kong's biotech, but time is critical. If Shanghai gets going, it could leave us behind," says Hong Kong University's Frederick Leung.

D.C.

development (R&D), it provides a genetically modified organism testing service, DNA fingerprinting for paternity cases and test kits for avian viruses.

But with few companies and no major research institutes, there are almost no opportunities for the graduates coming through the expanding biological-science university programmes. "There is a lot of talent, but few chances," says Ip, who has watched many of her highly qualified graduate students struggle to find research positions.

Some believe there should be a more concentrated use of resources to create large-scale research institutes. Advocates say that these would not only take in more students, but also use the resources more efficiently. Certainly the strong competition for funds contrast sharply with the streamlined approach to biomedicine in Singapore, where a centralized effort has led to large institutes and many job openings.

Many are pinning their hopes for Hong Kong's future on the ITC's latest endeavour, the Applied Science and Technology Research Institute (ASTRI). ASTRI's main function will be to provide 'midstream' research, bridging the gap between laboratories to commercial applications. "Hong Kong venture capital is less willing to invest in early-stage projects — it prefers to wait until a prototype is ready," says Simon Wong, who has just taken over as chief executive of ASTRI after ten years in Stanford's electrical engineering department. Wong hopes to bring some of the venture spirit he picked up in Silicon Valley to his native Hong Kong.

ASTRI aims to do R&D that will draw investment to startups. "At first this might be difficult, and we might have to make our own spin-off companies," says Wong. ASTRI plans to establish a cluster of R&D companies within the next ten years in the Hong Kong

Science Park, currently under construction.

Biotech will probably not be an early focus, but ASTRI will have one large biotech project under its wing from the beginning — the Institute of Chinese Medicine (ICM; see *Nature* **412**, 7; 2001). ASTRI aims to achieve the standardization and other R&D necessary to commercialize traditional Chinese medicine.

Hong Kong researchers have been working with traditional medicine for some time, but little has been achieved in turning this research into products. With great optimism for the new institute, endowed with US\$64 million, Ip anticipates that the BRI's "efforts in promoting Chinese medicine using modern molecular and cellular approaches have finally come to fruition".

For basic research, researchers can also apply for funding from the Research Grants Council. But these two-to-three-year grants are often found to be unsatisfactory. "They don't allow much flexibility. They should be five years. You often can't get results until your second year," says Walter Ho, a chemist at the Chinese University of Hong Kong. The short-term thinking that limits biotechnology investment can also take its toll on university fundamental research — and day-to-day life in general.

One researcher suggested reducing the number of grants, while increasing the quality. You just don't have enough time to cultivate young researchers, he says. The amount of time it takes the grants to get reviewed — nine months, compared with about six in the United States — also can hold up research plans, says Andrew Miller, a HKUST biologist from Scotland. "To be fair, though, the approval takes so long because of the difficulties required in getting a good external review," he adds.

But taking the good and the bad together, the final question is "can you do the work you want to do?" says Miller. "And the answer is 'yes, you can.'" Generous funding for equipment and a devoted technician for most researchers make for a productive working environment.

Conditions are not yet perfect, but Hong Kong is known as a place for betting. Most of its major biotech initiatives — including the BRI, the Hong Kong Institute of Biotechnology and the ICM — were funded by the Hong Kong Jockey Club, a government-approved gambling monopoly. For the time being, the postdocs, bioscientists and biotechnologists are willing to gamble — to give Hong Kong a few more years — as they wait to see the outcome of the ICM, ASTRI and other initiatives. ■

David Cyranoski is *Nature's* Asian-Pacific correspondent.

Applied Science and Technology Research Institute (ASTRI)

♦ <http://www.info.gov.hk/itc/eng/infrastructure/astri.shtml>



Nancy Ip and other researchers would like to see more long-term private investment in Hong Kong research.

Big salaries, little grants

Jobs at Hong Kong's universities are highly competitive, largely because of the pay. Salaries for full professors start at HK\$100,000 — a generous annual salary of about US\$150,000. "The standard of living is a major draw," says Andrew Miller, from the Hong Kong University of Science and Technology (HKUST).

The largesse extends to the lower levels as well, and most graduate students (HK\$15,500 per month), research associates (Hong Kong's version of postdocs) and

technicians (HK\$28,000 per month) are well-paid compared with their counterparts around the world.

But the inflated salaries can also be a problem — when you have to pay them. "The main drawback in Hong Kong is that there is not enough money for postdocs," says Daniel Chan, Chair of the Department of Zoology at the University of Hong Kong. According to one researcher: "After you pay a postdoc, there's little left for laboratory expenditures."

D.C.